2014

An Evaluation Of Sample Size Re-Estimation Adaptive Designs And Delayed-Start Designs For Alzheimer'S Disease Trials

Guoqiao Wang
University of Alabama at Birmingham

AN EVALUATION OF SAMPLE SIZE RE-ESTIMATION ADAPTIVE DESIGNS AND DELAYED-START DESIGNS FOR ALZHEIMER'S DISEASE TRIALS

by

GUOQIAO WANG

GARY R. CUTTER, COMMITTEE CHAIR
RICHARD E. KENNEDY
LON S. SCHNEIDER
ALFRED A. BARTOLUCCI
INMACULADA B. ABAN

A DISSERTATION

Submitted to the graduate faculty of The University of Alabama at Birmingham,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

2014

AN EVALUATION OF SAMPLE SIZE RE-ESTIMATION ADAPTIVE DESIGNS AND DELAYED-START DESIGNS FOR ALZHEIMER'S DISEASE TRIALS

GUOQIAO WANG

BIOSTATISTICS

ABSTRACT

The goal of this dissertation is to investigate the effect of novel clinical trial designs for Alzheimer's disease (AD), and to provide applications for their use in real trials. The data used for our simulation is a meta-data base of completed trials. In the first paper, we investigate the sample size re-estimation (SSR) adaptive design based on the effect size and the variance without taking into account the longitudinal feature of the trials. In the second paper, we take advantage of that feature to explore the SSR based on the variance of the rate of change in the longitudinal measurements. Finally, in the third paper, we extend the delayed-start (DS) design to AD by proposing some of the crucial design parameters. We also investigate the power of the DS design, and compare it to the power of the typical randomized parallel-group design.

Through our simulations, we discover that SSR based on the effect size or the variance without taking into account of the longitudinal feature of the trial can be effective for trials with small or moderate initial sample sizes. However, when the initial sample size is over 200, the gain in power after SSR is no longer significant. After incorporating the longitudinal feature, we show that SSR based on the rate of change is not only effective, but also allows the luxury to adapt the sample into two ways: increase the number of recruits or add the number of measurements. However, increasing the number of recruits is more likely. Finally, for the DS design, we prove that the optimal

sample size allocation ratio is 1:1:1; the optimal weight has a simple formula; the correlation between slopes can be negative and positive; and the optimal treatment-switch point is the middle point or the second one of the middle two.

Keywords: Alzheimer's disease, mild cognitive impairment, sample size re-estimation, adaptive design, longitudinal study, delayed-start design

DEDICATION

I dedicate my dissertation research to my wife, Jige Guo, my father, Zeyou Wang, my mother, Xiuge Han, and my sons, Joshua and Jeremy Wang. Without their support, motivation, encouragement, and dedication, this research would not have been possible.

ACKNOWLEDGEMENTS

I would like to express my most sincere and deepest gratitude towards my advisor, Dr. Gary R. Cutter for his excellent guidance, generous caring, endless patience, tireless encouragement, and providing me with a wonderful atmosphere for doing research. Moreover, Dr. Cutter also offers invaluable guidance and timely counseling in helping me to become a better husband and father. Without him, neither I, nor my family would where we are today.

I am very thankful for Dr. Richard Kennedy, for meeting with me every week for more than 2 years and providing all possible assistance in programming, theoretical derivation, drafting and so on. I must thank Dr. Lon Schneider for allowing me to use the meta-database, aiding my understanding in Alzheimer's disease, and guiding me in writing those papers. I also would like to thank Dr. Inmaculada Aban for her rigorous attitude toward my homework and research, being available whenever I need her help, and encouraging me many times. I definitely want to thank Dr. Alfred Bartolucci for offering alternative methods, providing feedback, and being patient.

I must thank Drs. Shan Zhao and James Wang from the University of Alabama for their generous help and invaluable suggestions on my way to changing my major. It is Dr. Zhao who first taught me how to write a computer programming and empowered me to overcome its intimidation. They are two of the kindest professors I ever had.

In addition, I am very thankful for my great friends Matthew Loop, Brandon J George, and Hwasoon Kim. We came into the department at the same year and very soon

bound together. They helped me to sort out my proposal and dissertation. Without them, I would not be where I am.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table *Page*

LITERATURE REVIEW

EFFECT OF SAMPLE SIZE RE-ESTIMATION IN ADAPTIVE CLINICAL TRIALS FOR ALZHEIMER'S DISEASE

EFFECT OF SAMPLE SIZE RE-ESTIMATION IN ADAPTIVE DESIGN CLINICAL TRIALS FOR ALZHEIMER'S DISEASE WHEN THE KEY RESPONSE IS THE RATE OF CHANGE

DESIGN PARAMETERS AND EFFECT OF THE DELAYED-START DESIGN FOR ALZHEIMER'S DISEASE

LIST OF FIGURES

Figures *Page*

LITERATURE REVIEW

EFFECT OF SAMPLE SIZE RE-ESTIMATION IN ADAPTIVE CLINICAL TRIALS FOR ALZHEIMER'S DISEASE

EFFECT OF SAMPLE SIZE RE-ESTIMATION IN ADAPTIVE DESIGN CLINICAL TRIALS FOR ALZHEIMER'S DISEASE WHEN THE KEY RESPONSE IS THE RATE OF CHANGE

DESIGN PARAMETERS AND EFFECT OF THE DELAYED-START DESIGN FOR ALZHEIMER'S DISEASE

Literature Review

Background and Motivation for Research

Mild Cognitive Impairment, Alzheimer's disease, and their negative impact

"AD is an age-related, progressive neurodegenerative disorder that gradually destroys a person's ability to remember, think, and even carry out the simplest tasks". AD is diagnosed mostly in people age 60 or older, and it is becoming increasingly common as the oldest "baby boomers" in U.S. turn 65 [1]. As of 2012, an estimated more than 5 million Americans are suffering from AD. The disease ravages the patients as well as the entire family, in emotional, physical, and financial ways. Currently, AD costs the health care system $200 billion a year [2]. The number of people affected in America by AD will jump to 13.5 million by 2050 [2]. Therefore, it is imperative that effective treatments be developed to prevent or delay the onset of AD, or to stop the progression of AD.

"Mild cognitive impairment (MCI) refers to the clinical condition between normal aging and AD in which persons experience memory loss to a greater extent than one would expect for age, yet they do not meet currently accepted criteria for clinically probable AD" [3]. It causes cognitive changes that may be noticed by the individuals experiencing them, but are not serious enough to affect daily activities. Although not all of those with MCI will develop Alzheimer's disease (AD) or another type of dementia, their risk is increasing [2].

Symptomatic and Disease-modifying Therapeutics for AD

There are two main AD therapeutics: symptomatic treatments and disease-modifying treatments. The former only mitigates the symptoms of AD, such as improvements on cognitive scores, relief in anxiety, and amelioration in low mood and irritability. The latter not only lightens the symptoms, but is also expected to delay the onset of the disability caused by AD or to slow down the progression of the disease course [4].

Available Therapeutics for AD, the Dominant Clinical Design Used in AD and Its Future Alternatives

Clinical trials for AD have been conducted for a little over 30 years. Prior to 1986, the methodology for conducting clinical trials in AD was virtually non-existent. In 1990, the US Food and Drug Administration (FDA) established guidelines for anti-AD drug trials. Under the 1990 guidelines, several drugs have been approved for symptomatic therapies [5, 6], including tacrine, donepezil, rivastigmine, galantamine, and memantine. But still no disease-modifying drugs are available. Currently, there are approximately 80 drugs for AD being investigated in over 200 clinical trials mostly phase I/II [7]. As part of the consequence of investigations, the randomized, double-blind, placebo-controlled, parallel-group clinical trial design has become the standard design according to FDA guidelines [8]. During the 1990's, a trial of 6 months duration with about 100-120 subjects per arm was generally considered sufficient to detect the treatment effects. However, due to the almost uniformly negative results of the initial clinical trials, sample

sizes have increased remarkably for both phase II (40 to 200 per arm) and phase III (201 to 842 per arm) and the trial duration has extended from 6 months up to 24 months[5]. Despite the enormous increase in the trial duration and the sample size, the lack of success in detecting an effective treatment using the typical design remains. Moreover, there has not any approved disease-modifying therapeutics. Potential causes of these negative trials included the lack of efficacy in the treatments, insensitivity of the primary outcome to cognitive changes, underpowered trials due to the inaccurate pre-trial estimates of treatment effect, and so on. Therefore, clinical trial designs which allow interim analyses and resultant modification of the ongoing trial to increase or adjust power, such as adaptive designs, have been recommended as alternatives[9]. One such adaptive design is the SSR adaptive design, which allows sample size adjustment based on the comparison between the interim treatment effect (or the interim variance) and the pre-trial treatment effect (or the pre-trial variance) [10]. However, both the typical design and the adaptive design are parallel group designs, which may not be able to distinguish the disease-modifying therapeutics from the symptomatic ones [11]. In order to facilitate the detection of disease-modifying treatments, the delayed-start (DS) design has been proposed. In the DS design, patients are randomly assigned to placebo or treatment for a pre-specified frame of time and then those (or a randomized portion of those) in the placebo group are also given the treatment. If patients who are on the treatment from the beginning of the study show similar effects as those who received the treatment later, the treatment effect if any is considered symptomatic, but not disease-modifying [4, 11]. These two novel designs have many advantages over the typical design. For example, the SSR adaptive design has the flexibility to adjust the trial for better efficacy, minimize the

number of patients exposed to the inferior treatment, avoids the long-term trials for drugs with limited efficacy, and better utilizes the most recent external or internal information in the ongoing trial. The DS design has the unique ability to declare the disease-modifying effect. However, there are also concerns when employing SSR, such as the reliability in estimating the overall treatment effect based on a relatively small interim sample (or, for longitudinal trials, the precision in predicting the final treatment effect using only the early measurements), and the tradeoff between the gain in estimated power versus the burden to recruit more subjects. The former concern is particularly relevant for clinical trials in AD, as heterogeneity in the course of the disease may introduce significant inaccuracies in estimating the final treatment effect based on interim analyses [12].

On the other hand, due to the complexity in determining the crucial design parameters such as the sample size allocation ratio in different treatment arms, the optimal time of treatment switch, the length of the before-treatment-switch period, the test statistic, and the power for given sample sizes; the DS design has not been successfully applied in any AD clinical trial to detect disease modifying treatments [6]. Therefore, this paper addresses the practical necessity to investigate the applicability of these novel designs before their implementation.

Clinical Design Background and Statistical Methods

A Brief Review of the SSR Adaptive Design

The concept of adaptive designs was first introduced into statistical community in 1978 [13], since then it has accumulated increasing interests. An adaptive design refers to

a clinical trial design that uses accumulating data at the interim analysis to modify certain aspects of the trial as it is ongoing without undermining the validity and integrity of the trial, and these adaptations are pre-determined rather than impromptu in order to avoid bias [14]. Therefore, adaptive designs offer the flexibility to learn from the accumulating data and apply what is learned quickly to an ongoing trial.

Compared to the typical parallel group design, the adaptive design is advantageous in several aspects: 1) it offers the flexibility to learn from the accumulating data and apply what is learned quickly to an ongoing trial; 2) it can improve efficiency either by reducing the number of patients exposed to the treatment with limited efficacy or by stopping the trial earlier for futility; 3) it reflects medical practice in the real world in that we want to learn from an ongoing trial and then use what we learned to improve it; and 4) it is proved efficient in early or late phase of clinical development [15]. As a consequence of rapid development in adaptive design methodology, adaptive designs now include a general set of methods such as adaptive randomization; adaptive dose-finding studies, seamless phase II/III designs, and sample size re-estimation (SSR), etc. In this dissertation, we focused on the SSR adaptive design.

The typical parallel design starts the trial with a pre-specified sample size, and modification regarding the sample size would not be allowed after the trial has started. In the absence of dropouts, the trial would end with the same sample size as specified at the beginning. The SSR adaptive design starts with a typical parallel design, and then allows the sample size to increase when the pre-trial treatment effect size was overestimated or the pre-trial variance of the outcome was underestimated, leading to a trial that concludes using a larger sample size retaining the more power specified at the beginning; and

allows early stopping or an overall decrease in the sample size when the pre-trial treatment effect size was underestimated or the pre-trial variance was overestimated, leading to a trial with the pre-specified power, but a smaller sample size (Figure 1).

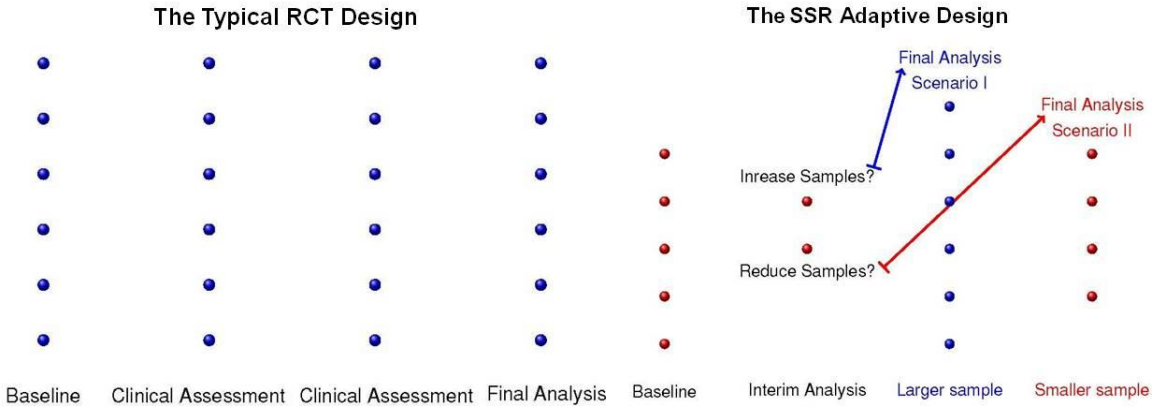

Figure 1. The simplified comparison between the typical clinical design and the SSR adaptive design: the former requires a fixed sample size; whereas the latter allows sample size adjustments

As noted previously, the SSR adaptive design offers some significant advantages over the typical parallel-group design including minimizing the number of patients exposed to potential toxic or inferior drugs, avoiding underpowered trials by adjusting the sample size, stopping the trials earlier for futility, and incorporating the most recent internal or external information into an ongoing trial [14]. All these unique features of the SSR adaptive design are potentially beneficial to clinical trials for AD. For example, the heterogeneity in the course of the disease leads to patients' inconsistent response over different treatments, thus information obtained from previous trials might not accurately reflect what would happen in the next trial, resulting in inaccurate estimates of the treatment effect or the sample size; so it is helpful to conduct SSR in order to right-size the trial to demonstrate efficacy or to right-end the trial to enhance efficiency.

When planning a trial, at least two design parameters are required to appropriately power a trial: the treatment effect to detect and the nuisance parameter related to the primary outcome such as its variance. Consequently, misspecification of these two design parameters may lead to overpowered, expensive, and lengthy trials or underpowered and inefficient trials. Correspondingly, a SSR can be conducted at the interim analysis to re-evaluate either or both of these two design parameters, thus generally speaking leads to two types of SSR: 1) SSR based on the treatment effect or based on a combination of the treatment effect and the nuisance parameter, which inflates type I error and thus requires corresponding adjustments; and 2) SSR based on the nuisance parameter only, which does not inflate type I error and thus requires no adjustments [16-19]. Regardless of the SSR method, a typical SSR adaptive design involves several steps: 1) obtaining pre-trial estimates design parameters such as the variance of the primary outcome, the treatment effect, and the sample size for beginning the trial; 2) performing an interim analysis at a specified time point re-estimates all or some of the design parameters; 3) based on the comparison of the re-estimated design parameter with the pre-trial ones, invoking a decision rule about the next phase of the trial such as increase or decrease the sample size; 4) analyzing the final trial results without inflating the type I error.

At the interim analysis, the design parameters can be re-estimated in a blinded fashion, meaning the patients' treatment assignment is unknown to trial personnel such as clinicians and statisticians; or in an unblinded fashion, meaning the treatment assignment is known to some of the trial personnel. The blinded SSR is usually preferable, especially for the SSR based on variances for a continuous outcome [14]. The blinded SSR is superior to the unblinded for reasons including: 1) it generally behaves as well as the

unblinded [17]; 2) it tends to minimize the inflation of type I error if any, particularly for moderate or large sample sizes [20]; 3) it is less likely to induce biases and to undermine the integrity of the trial [14]; and finally, 4) it is preferred from a regulatory standpoint [16]. In this study, both the blinded SSR based on the variance of the primary outcome and the unblinded SSR based on the effect size were investigated. When conducting a blinded SSR based on the variance, the variance needs to be estimated blindly at the interim analysis, and the estimate can be done by two main methods: the expectation–maximization (EM) algorithm [21] and "the pooled sample variance with adjustment based on the difference between the means presumed in the alternative hypothesis" [19]. A detailed comparison between the two methods will be presented later.

Comparison between the SSR adaptive design and the group sequential design in the context of AD clinical trials

From the standpoint of frequentist statistics, two main designs provide the luxury to adjust the ongoing trials based on the accumulating data: the group sequential design (GSD) and the adaptive design [22]. The GSD samples groups of observations for interim analyses, and consequently, the trial can be stopped at any interim analysis and after any of these groups for safety, efficacy/futility or both [15]. The GSD was introduced in 1947, and has been well-developed and well-accepted [23]. Additionally, GSDs were argued to perform more efficiently than adaptive designs in certain circumstances such as when the key parameters are known in the beginning of the clinical trials [24]. Therefore, we make the following comparison between the two methods in the context of AD to justify our choice of the SSR adaptive design over the GSD.

1) *The main objective*

The two types of designs serve fundamentally different purposes. The GSD aims for early stopping, whereas, the SSR adaptive design aims to increase the sample size flexibly to ensure the study power. The failed 18- and 24-month trials in AD have showed that early stopping is not realistic; instead, increasing the power is the primary purpose of using novel clinical trial designs, which means the SSR adaptive design fits the goal better.

2) *The time and number of interim analyses*

The GSD usually involves more than two interim analyses, and thus requires earlier first interim analysis than the adaptive design, which, in most situations, involves only one interim analysis. In addition, AD is chronic and progresses slowly, thus an early interim analysis may not reveal useful information or provide accurate estimates.

3) *The sample size in the beginning of the study*

The GSD generally "starts large, and if you can, stops early", meaning GSDs usually start with a relatively large sample size, sometimes even with a maximal sample size evaluated using the smallest treatment effect. On the other hand, the SSR adaptive design "starts small and asks for more if necessary", meaning it usually starts with a relatively small sample size, and then uses the accumulated data at the interim analysis to decide whether or not to increase the sample size [25]. Many failed clinical trials with relatively large sample size in AD have indicated that the "start large"

strategy probably is not the optimal one. Furthermore, given the small effect sizes typically observed in AD, the "start large" method would likely yield such a large sample size that it could not realistically be implemented.

4) *The impact of the length of the recruiting time*

The recruiting time for AD trials usually is not long relative to the trial duration, thus patients' outcome measurements will cluster in a very short time interval, which means that to sample groups of measurements orderly with pre-determined time space when all of them are available is inefficient and even unethical.

On the contrary, the clustered measurements serve the SSR adaptive design very well since the study starts with a small sample size and data are intended to be collected as much as possible within a short period of time for the purpose of reliable estimates.

5) *The flexibility in the sample size*

The GSD, like the traditional randomized, placebo-controlled, parallel-group design, generally requires a fixed sample size from the beginning to the end of the trial, whereas, the SSR adaptive design allows the interim analysis to determine the final sample size, and thus produces a flexible maximum sample size. This flexible maximum sample size is a helpful and new paradigm, and enables the SSR adaptive design to have larger power than the GSD. On the other hand, due to the multiple interim analyses, the GSD usually results in loss of power [26].

6) *The preferable study object*

Studies with GSDs almost always involve mortality or irreversible morbidity as primary efficacy endpoints (thus ethics require possible early termination), while the SSR adaptive design is often used for trials with non-life-threatening or chronic diseases with continuous or binary outcomes [26]. The chronic nature of AD along with Alzheimer's disease assessment scale cognitive sub-scale (ADAS-Cog) as the primary continuous outcome certainly fits the frame of the latter properly.

7) *The type of sample adjustment*

For clinical trials with longitudinal data, the sample can be adjusted in two ways: 1) recruit more subjects while retain the number of longitudinal measurements, meaning extend the duration of the trial; and 2) increase the number of longitudinal measurements while retain the number of subjects. The GSD generally requires the duration of the trial to be pre-determined in order to schedule multiple interim analyses in the design stage, and thus it limits the sample adjustment to the number of subjects only.

In sum, the SSR adaptive design is a superior option for AD over the GSD.

Methods for blinded estimate of the variance

A SSR adaptive design starts with a typical parallel-group design: the treatment group and the placebo group. At the interim analysis, outcomes from both groups are

collected for the estimate of the common variance without knowing the groups they belong to.

Two methods for the blinded estimate of the interim variance were proposed by Gould and Shih [21]: "the pooled sample variance with adjustment based on the treatment effect presumed under the alternative hypothesis" (henceforth, referred as the pooled-sample-variance method); and the EM algorithm which is independent of the presumed treatment effect. Govindarajulu extended the pooled-sample-variance method to outcomes from arbitrary distributions [19]. Gould and Shin claimed that the maximum likelihood estimate (MLE) of the common variance by the EM algorithm preserves the blindness and is very satisfactory. However, Friede and Kieser showed that the EM algorithm: 1) depends on the initial values; 2) its stopping rule may not be able to guarantee the convergence of the MLE to the true variance; 3) it is only appropriate for trials with the simple randomized assignment, e.g. 1:1 sample allocation ratio [27]. Waksman improved Gould and Shin's procedure and his updated version overcomes the aforementioned flaws; however, the new procedure still leads to a negatively biased estimate with a large standard deviation. Additionally, he concluded that "when the standardized treatment effect is 1 or less, which is typical in most trials", the pooled-sample-variance yields a better estimate despite of its positive bias [28]. We extended Waksman's work and investigated the impact of skewness on the estimate for both methods.

The non-normal data with given skewness and kurtosis were generated by Fleishman's polynomial method [29, 30]. Assume random variance $X \sim N(0, 1)$, then let

$$Y = a + bX + cX^2 + dX^3,$$

with $E(Y) = 0, E(Y^2) = 0, E(Y^3) = \gamma_1$, and $E(Y^4) = \gamma_2 + 3$, where γ_1 and γ_2 are the pre-specified values of skewness and kurtosis. Then the corresponding coefficients satisfy the following equations.

$$a = -c$$

$$b^2 + 6bd + 2c^2 + 15d^2 - 1 = 0$$

$$2c(b^2 + 24bd + 105d^2 + 2) - \gamma_1 = 0$$

$$24\{bd + c^2(1 + b^2 + 28bd) + d^2(12 + 48bd + 141c^2 + 225d^2)\} - \gamma_2 = 0$$

Fleishman solved the 3rd equation for c, and then substituted the resultant expression into the 2nd and the 4th equations. That yielded two equations for solving two variables. Fleishman used a modified Newton method to accomplish this. We applied the same method and obtained a list of the coefficients with corresponding values of skewness and kurtosis (Table 1).

Table 1. The combinations of skewness, kurtosis, and corresponding coefficients

Combination	Skewness	Kurtosis	b	c	d
1	0.5	0.5	0.97343	0.08045	0.006647
2	0.6	0.5	0.98755	0.10020	0.000784
3	0.7	0.5	1.00633	0.12311	-0.007253
4	0.8	0.5	1.03156	0.15154	-0.018621
5	0.5	0.6	0.96255	0.07872	0.010305
6	0.6	0.6	0.97572	0.09777	0.004867
7	0.7	0.6	0.99310	0.11957	-0.002505
8	0.8	0.6	1.01622	0.14606	-0.012761
9	-0.5	0.5	0.97343	-0.08045	0.006647
10	-0.6	0.5	0.98755	-0.10020	0.000784
11	-0.7	0.5	1.00633	-0.12311	-0.007253
12	-0.8	0.5	1.03156	-0.15154	-0.018621

Combination	Skewness	Kurtosis	b	c	d
13	-0.5	0.6	0.96255	-0.07872	0.010305
14	-0.6	0.6	0.97572	-0.09777	0.004867
15	-0.7	0.6	0.99310	-0.11957	-0.002505
16	-0.8	0.6	1.01622	-0.14606	-0.012761
17	1.0	0.5	1.11465	0.25852	-0.066013
18	1.2	0.5	-0.91451	0.15709	-0.024614
19	-1.0	0.5	1.11465	-0.25852	-0.066013
20	-1.2	0.5	-0.91451	-0.15709	-0.024614
21	0.4	1.0	0.91613	0.05736	0.026170
22	0.6	1.2	0.91664	0.08706	0.024639
23	0.8	1.4	0.92471	0.11985	0.019872
24	1.0	1.6	0.94243	0.15938	0.010500
25	1.2	1.8	0.97458	0.21578	-0.007486
26	-0.4	1.0	0.91613	-0.05736	0.026170
27	-0.6	1.2	0.91664	-0.08706	0.024639
28	-0.8	1.4	0.92471	-0.11985	0.019872
29	-1.0	1.6	0.94243	-0.15938	0.010500
30	-1.2	1.8	0.97458	-0.21578	-0.007486

Based on these lists, we generated non-normally distributed data based on $N(0,1)$ and $N(0.25,1)$, $N(0,2)$ and $N(0.5,2)$, $N(0,3)$ and $N(0.75,3)$, $N(0,4)$ and $N(1.0,4)$, and finally $N(0,5)$ and $N(1.25,5)$, and then estimated the common variance using the EM algorithm and the pooled-sample-variance method for different sample sizes per group. Our simulation showed that the EM algorithm is more vulnerable to skewness and kurtosis (Figure 2).

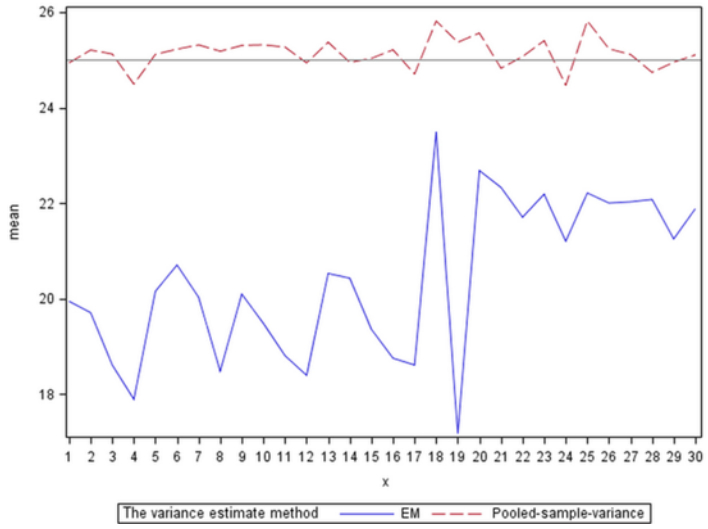

Figure 2. The impact of skewness on the blinded estimate of variances by estimate methods based on $N(0,5)$ and $N(1.25,5)$.

All things considered, in this study, the pooled-sample-variance method is chosen for the blinded estimate of the interim variance.

Real patient data used in the simulation

Participants for the simulations were drawn from a meta-database of clinical trials and observational studies [31]. Of all the studies, 8 of them were used for this dissertation (Table 2).

Table 2. Studies used in this dissertation

Study (code)	Design	N	Duration (months)
Selegiline, vitamin E (SL)	RCT, moderate to severe AD	341	24
Prednisone (PR)	RCT, mild to moderate AD	138	16
Conjugated estrogens (CE)	RCT, mild to moderate AD	120	15
Memory impairment study (MIS)	RCT, MCI	769	36
Simvastatin (LL)	RCT, mild to moderate AD	406	18
Vitamins B (HC)	RCT, mild to moderate AD	409	18
DHA (DHA)	RCT, mild to moderate AD	402	18
ADNI (ADNI)	Observational, AD, MCI, normal	800	36 (AD) 48 (MCI) 48 (NL)

Abbreviations: RCT, randomized clinical trial; LL, lipid lowering; HC, homocysteine; DHA, Docosahexaenoic Acid; ADNI, Alzheimer's Disease Neuroimaging Initiative; NL, normal.

Research Goals

Paper 1

We begin with the SSR based on the effect size and the variance using only a single measurement. We evaluate the impact of SSR on power and final sample sizes. We also consider other factors that potentially affect the behavior of SSR, such as the time of SSR, the initial sample size, and the duration of the trial.

Paper 2

Taking advantage of the longitudinal data, we evaluate the SSR method based on the variance of the rate of change of the longitudinal data. This method leads to adjustments in the final sample size or in the total number of measurements for each

subject. We examine the differences between these two types of adjustments and potential factors that may affect them.

Paper 3

First, we improve and propose the values of the crucial design parameters in DS design. Second, we extend the assumption of the variances. Finally, through simulation, we compare the power of the DS design with the typical randomized parallel-group design, and evaluate the impact of the variance assumption on power.

Future Research

We conclude this dissertation with a discussion on limitations of our current work and directions for future research.

EFFECT OF SAMPLE SIZE RE-ESTIMATION IN ADAPTIVE CLINICAL TRIALS FOR ALZHEIMER'S DISEASE

GUOQIAO WANG, RICHARD E. KENNEDY, GARY R. CUTTER, LON S. SCHNEIDER

In preparation for submission

Format adapted for dissertation

1 INTRODUCTION

The number of individuals with AD continues to grow worldwide with the aging of the population [1]. Although a handful of modestly effective symptomatic treatments have been developed using the typical randomized clinical trial (RCT) design, clinical trials to identify effective disease-modifying treatments to slow the progression of AD have been uniformly negative [2-4]. There are several potential causes of these negative trials, including the lack of efficacy in the treatments, insensitivity of the primary outcome to treatment changes, and low power due to the inaccurate pre-trial estimates of the treatment effect. Therefore, clinical trial designs which allow interim analyses and resultant modification of the ongoing trial to increase power (adaptive designs) have been recommended [5]. One such approach is the sample size re-estimation (SSR) adaptive design, which allows sample size adjustment based on the comparison between the interim treatment effect (or the interim variance) to the pre-trial treatment effect (or the interim variance) [6].

A simplified comparison between the typical RCT design used in AD and the SSR adaptive design is illustrated in Figure 1. The typical RCT design starts the trial with a pre-specified sample size, and modification regarding sample size would not be allowed after the trial has started. In the absence of dropouts, the trial would end with the same sample size as specified at the beginning. The SSR adaptive design allows the sample size to increase when the pre-trial treatment effect size was overestimated or the pre-trial

variance of the outcome was underestimated, leading to a trial that concludes using a larger sample size to retain the power specified at the beginning. It also allows early stopping or an overall decrease in the sample size when the pre-trial treatment effect size was underestimated or the pre-trial variance was overestimated, leading to a trial with the pre-specified power, but a smaller sample size. This flexibility can not only adjust the trial to improve efficacy, but also provide other advantages over the typical RCT design, such as minimizing the number of patients exposed to inferior treatment, avoiding long-term trials for drugs with limited efficacy, and better utilizing the most recent external or internal information of the ongoing trial. However, there are potential concerns when employing SSR, such as the reliability in estimating the overall treatment effect based on a relatively small interim sample (or, for longitudinal trials, the precision in predicting the final treatment effect using only the early measurements), and the tradeoff between the gain in power versus the burden to recruit more subjects. The former concern is particularly relevant for clinical trials in AD, as heterogeneity in the course of the disease may introduce significant inaccuracies in estimating the final treatment effect based on interim analyses. This study was designed to use simulations based on real patient data to investigate the behavior of SSR in an adaptive trial design for AD.

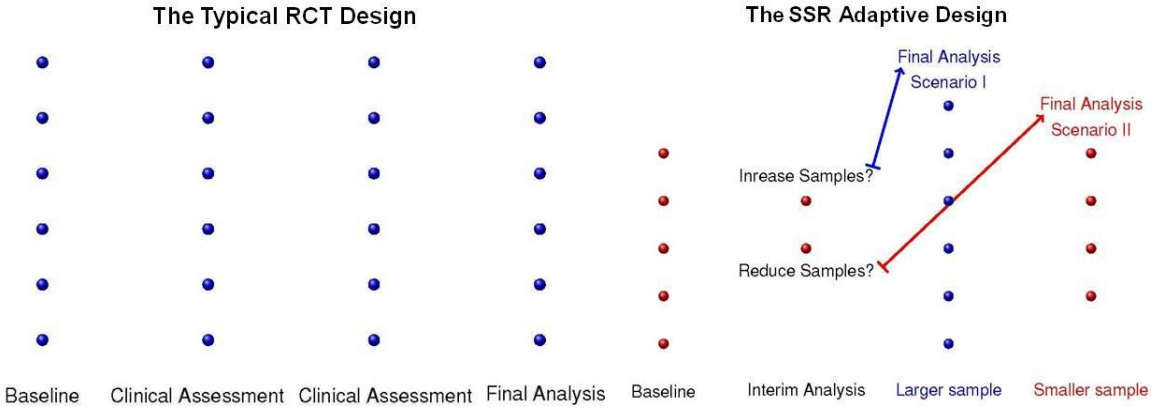

Figure 1. The simplified comparison between the typical clinical design and the SSR adaptive design: the former requires a fixed sample size; whereas the latter allows sample size adjustments

2 METHODS

2.1 Study Overview

Participants for the simulations were drawn a meta-database of 9 clinical trials and observational studies [7]. The primary outcome was the ADAS-Cog, which evaluates memory, reasoning, orientation, praxis, language, and word finding difficulty, and is scored from 0 to 70 errors, with higher scores indicating greater impairment [8]. Clinical assessments were done at 6-month intervals over the first 2 years.

2.2 Simulation Methods

Simulations were conducted under a detailed protocol [9], similar to our previously published approach [7, 10], to reflect clinical trials for an experimental drug for AD or MCI with one treatment group and one placebo group, 1:1 allocation ratio, and parameters for the distribution of ADAS-Cog selected to be consistent with previously published trials and ADNI [11, 12]. Clinical trials with sample sizes of 50, 100, 200, 300, and 400 per group, trial durations of 12 months or 18 months for AD and of 18 months or 24 months for MCI, and dropout rates of 20% or 40% in both groups, were simulated. For each scenario, a separate set of patients was constructed by randomly choosing from the meta-database with replacement, i.e., patients from the dataset could be present in the simulated groups more than once in the same or different treatment groups. The placebo group outcome was the score for the subject at the specified time point in the meta-database, with normally distributed random error with mean 0 and standard deviation 1

added to minimize ties in the outcome. For each subject in the treatment group, effect sizes of 0.15 and 0.25 (representing treatment effects of small to medium size) were used to compute simulated treatment results. The individual treatment effect was randomly generated from a χ^2 distribution with mean equal to the expected treatment effect (effect size times the pooled group standard deviation) to allow for a more realistic distribution of declines over time, where a few patients may fail or worsen more markedly because of the skewness of the χ^2 distribution than would be predicted by a normal distribution. As successful treatments would lead to smaller increases on the ADAS-cog than placebo, the individual treatment effect was shifted by subtracting two times the expected treatment effect, then adding the result to the patient's score at the specified time point in the database. For example, if a is the ADAS-Cog score at a given time point in the database, then $a + \chi_z^2 - 2 * z$, is the corresponding score in the simulated treatment group, where $z = effect\ size * sd$ and sd is the sample standard deviation of the change in ADAS-Cog from baseline. Assume $a = 24$, effect size is 0.25, sd is 8, and the randomly generated treatment effect from the χ_z^2 is 3, then the ADAS-Cog score used in the simulation would be 23. While a patient may be reused in the analysis, the actual value used would be modified by this randomly selected amount, hence making it slightly different.

2.3 Time Points Used for SSR

For a typical RCT AD, patients' enrollment times vary (Figure 2), leading to different number of available measurements for each patient at the interim analysis. In this example, at 12 months, patient 1 had 3 measurements available; patients 2 to 5 had 2; while patient 6 had only 1. In this paper, for a given trial with an initial sample size of 50

per arm, 'SSR at 12 months' means that all the patients enrolled and had been measured for up to at least 12 months. Notably, some patients were followed longer and their measurements were truncated at 12 months implying each patient would have the opportunity for 3 measurements. This truncation leads to power loss, but depends heavily on the recruitment rate. Thus for simplicity, we truncated the follow-up at this point so our results were not strictly dependent on the recruitment rate.

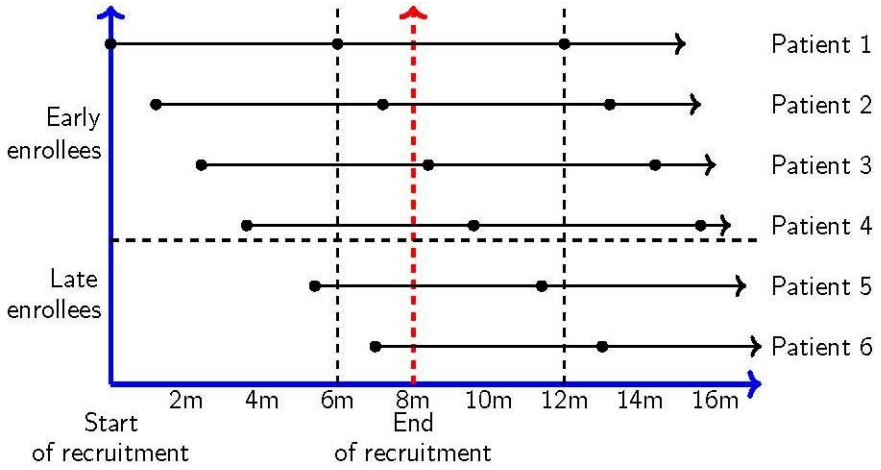

Figure 2. The enrollment times vary in a trial. The number of measurements per patient varies depending on enrollment times and the time of SSR performed. For example, at 12 months, patient 1 has 3 measurements; patients 2 to 5 have 2; and patient 6 has only 1.

2.4 Estimation Methods Used for SSR

SSR based on interim variances (henceforth, referred as "variance only method") and SSR based on interim effect sizes (henceforth, referred as "effect size method") were used, and both methods assumed equal variances in the treatment group and the placebo group. The "variance only method" assumes that the pre-trial estimate of the mean difference between treatment and placebo groups is accurate, and only the variance is uncertain, thus needs re-estimation. At the interim analysis, the variance of ADAS-cog

was estimated and compared with the pre-trial estimate, and then the sample size was adjusted based on the following equation:

$$N = \frac{\hat{\sigma}_i^2}{\hat{\sigma}_0^2} N_0.$$

Where, N is the re-estimated sample size, N_0 is the initial sample size, and $\hat{\sigma}_i^2$ and $\hat{\sigma}_0^2$ are the interim and the estimated pre-trial variances of the outcome, respectively. In our analysis, $\hat{\sigma}_i^2$ was estimated using pooled data (to mimic blinding to treatment in a clinical trial) as $\hat{\sigma}_i^2 = (N_i - 1)/(N_i - 2)(S^2 - \Delta^2/4)$, where N_i is the total sample size at the interim analysis, S^2 is the pooled sample variance, and Δ is the pre-trial estimate of the treatment effect [13]. This method does not inflate type I error, thus no adjustment to the α level is required.

The "effect size method" assumed that the estimate of the pre-trial estimate of the mean difference between treatment and placebo groups, as well as the pre-trial variance, is uncertain. At the interim analysis, both would be re-estimated and the initial sample size was adjusted based on the formula given by Chang [14]:

$$N = \left|\frac{E_0}{E_i}\right|^a N_0,$$

where, E_0 and E_i are the pre-trial and the interim observed effect sizes, and a is a tuning parameter and that is often chosen to be 2 because of the squared relation between the sample size and the effect size. E_i was approximated as $E_i = \Delta_i/S$, where, Δ_i is the observed treatment difference at the interim analysis, and S is the pooled sample deviation. This method requires unblinding of the treatment code, which must be

monitored carefully and kept to a minimum of individuals to preserve trial integrity. In addition, it does not preserve the type I error, so adjustment to the α level is required.

The pre-trial variances of ADAS-Cog scores for MCI and AD trials used in this study were 16 and 64 respectively, which were conservatively estimated based on the placebo outcomes of previous trials [3]. A single SSR was conducted at 6 months and 12 months. Increases in sample size are not necessary if significance of the treatment difference is achieved at the interim analysis, or if the treatment effect is as large as or larger than that hypothesized a priori, or if the variance is as small as or smaller than that hypothesized. For both methods, we assumed restricted designs [15], which means the initial sample size may be increased but not decreased. The latter restriction was a practical consideration, since in many chronic conditions; recruitment is often completed by the time of SSR.

2.5 Statistical Analysis

The primary analysis method was the Wilcoxon test of differences in ADAS-cog between the treatment group and the placebo group; missing values were imputed using last observation carry forward (LOCF) because of its simplicity and the assumption of non differential dropout as well as the longitudinal nature of the data [16]. The secondary analysis method was a mixed effects linear model, which tested the difference in the slopes of the ADAS-cog between the treatment group and the placebo group. For all analyses, the missing data pattern present in the meta-database was used to realistically simulate dropouts. Observations were missing in simulated datasets in cases where they were originally missing in the meta-database. Because of our use of treatment effect

applied to selected samples, differential dropout caused by informative censoring was not included into the comparison.

One thousand simulations were carried out for each scenario so that estimates of power could be obtained up to three digits. Power is defined as the proportion of 1000 simulated trials per scenario with p values less than or equal to 0.05. All analyses were performed using SAS software, Version 9.2 (SAS Institute, 2008).

3 RESULTS

SSR at 6 months resulted in highly variable outcomes for both sample size increases and power improvement regardless of SSR method (Figures 3 and 4). Approximately 25% of trials required at least a doubling of the sample size regardless of initial sample sizes. When the initial sample size per treatment group was 50, half of SSR projected no increase in sample sizes. After SSR, the gain in power varied by initial sample sizes, trial durations, and effect sizes. For example, given an MCI trial with effect size 0.25, duration of 18 months, and SSR at 6 months based on variances, the power of the trial on average increased from 38.8% to 61.3% for initial sample sizes of 50 per group and from 64.7% to 88.1% for initial sample sizes of 100 per group. In contrast, the gain in power is less dramatic for an AD trial under the same setting, e.g. the power on average increased only from 30.8% to 42.2% for initial sample sizes of 50 per group and from 53.0% to 69.4% for initial sample sizes of 100 per group. However, when the initial sample size was over 200, the gain in power was negligible regardless of the type of trials. When the effect size was smaller, the power before SSR as well as after also became smaller; however the gain in power actually increased over larger initial sample sizes

(Figure 3). Under the same SSR method, the longer trial duration did not generate larger gain in power (Figure 5). In contrast, SSR at 12 months showed greater gains in power, but were still highly variable, ranging from 0% to 44%, with no clear increase in power over larger initial sample sizes (Figure 6).

The "effect size method" generally resulted in greater gain in power than the "variance only method" (Figure 4). However, the greater gain was at the price of larger increase in the re-estimated sample sizes (Table 1), and it diminished over larger initial sample sizes. The two SSR methods generated very similar results for both AD and MCI clinical trials. On average, both the gain in power and the increase in sample sizes after SSR are slightly larger for Wilcoxon tests than for the mixed effects linear model tests.

Table 1. Increase in sample sizes after SSR by initial sample sizes and by SSR methods.

SSR method	Initial Sample Sizes				
	50	100	200	300	400
	Increase in sample sizes after SSR (mean(std))				
SSR based on variances	43(18)	85(25)	170(35)	253(43)	338(50)
SSR based on effect sizes	166(219)	210(226)	272(244)	303(253)	341(259)

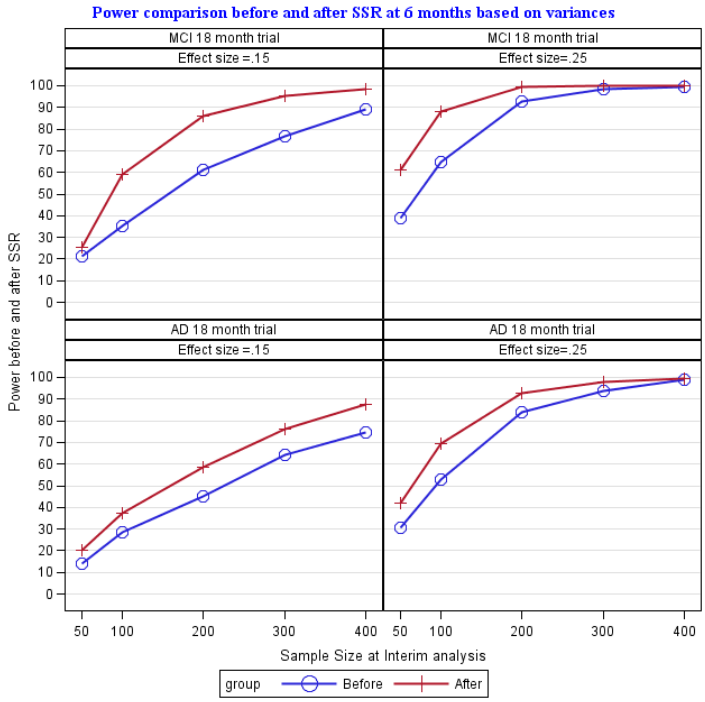

Figure 3. Power comparison before and after SSR based on variances at 6 months

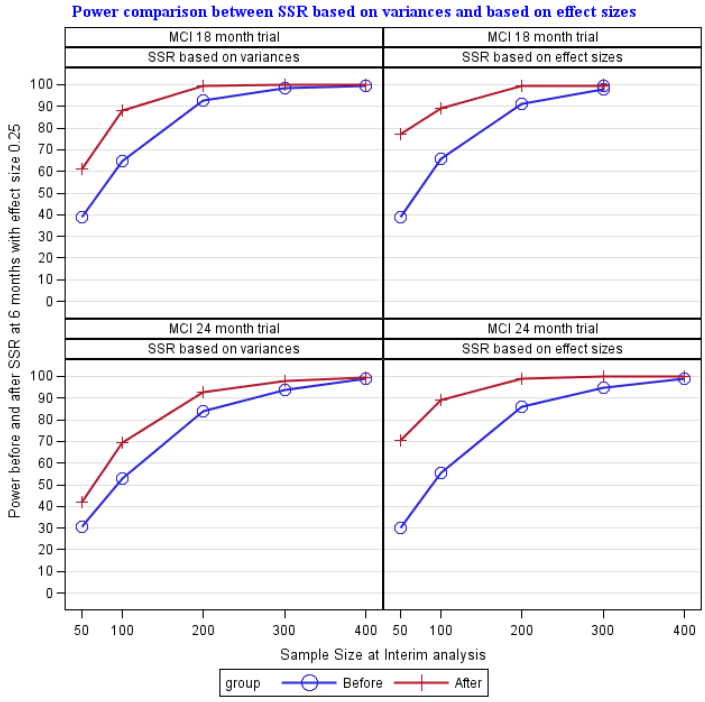

Figure 4. Comparison between SSR at 6 months based on variances and based on effect sizes

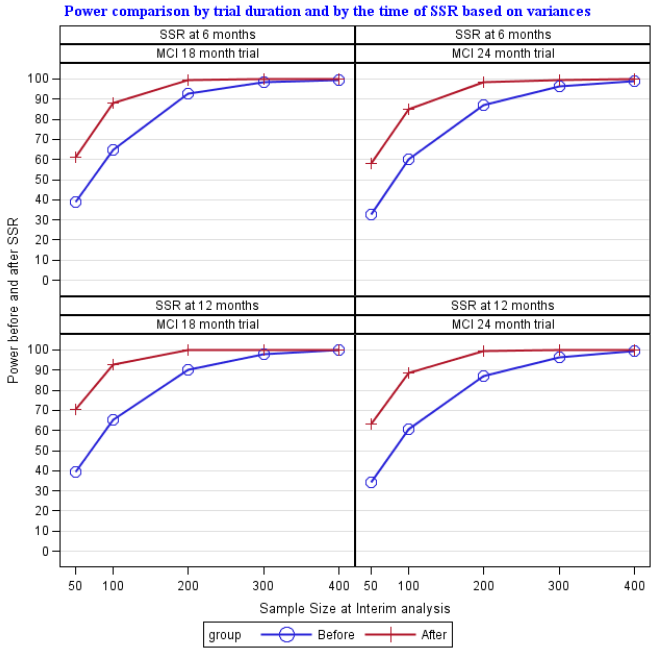

Figure 5. Power comparison by trial durations and by the time of SSR based on variances.

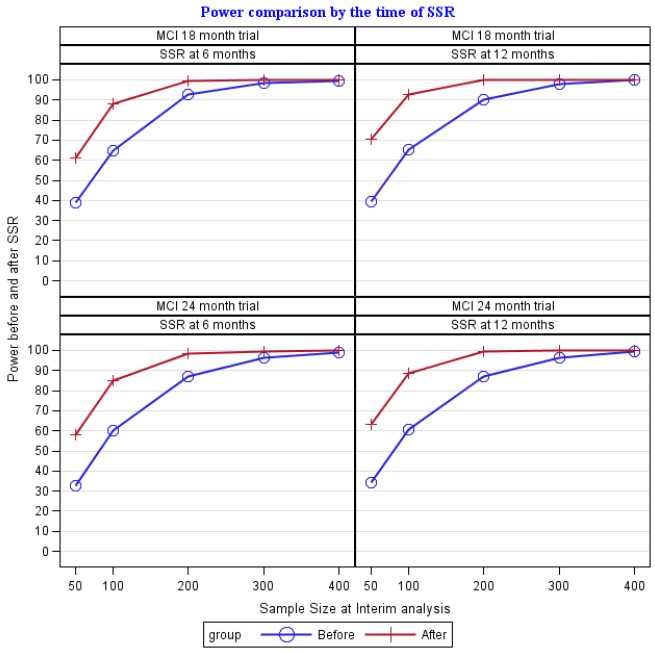

Figure 6. Power comparison by the time of SSR.

4 DISCUSSIONS

Based on our simulation, the SSR adaptive design can be effective for clinical trials in AD and MCI under certain circumstances. However, the effectiveness depends on several factors, such as the number of subjects accumulated for the interim analysis, the true treatment effect size, and the type of uncertainty in the pre-trial estimates (effect sizes or variances). Too few subjects accumulated for the interim analysis might lead to imprecise estimates of the treatment effect or the variance, thus resulting in poor prediction of sample size adjustments; with too many subjects, subjects are already enrolled and the trial without SSR already has adequate power. The smaller the true treatment effect, the more subjects are needed at the interim analysis in order to obtain precise estimates. Although the uncertainty in the pre-trial estimates determines the SSR method, the "variance only" method would be preferred over the "effect size" method [17, 18] and emphasizes the importance of pre-trial estimates of the difference between treatment and placebo groups. Based on our simulation, the former on average resulted in less gain in power than the latter; however, the latter tends to overshoot the final sample size, leading to recruitment of a much larger number of subjects than necessary.

For a longitudinal study, longer trials lead to more power for an effective treatment. However, our simulation indicated little difference in power between 18 months and 24 months trials after SSR. One explanation is that the relatively small treatment effect was not enough to overcome the heterogeneity and inconsistency in ADAS-Cog within a 6-month frame of time. This would also explain the lack of differences between SSR at 6 months and 12 months. An alternative would be to measure more frequently, e.g. every 3 months, and use more measurements at the interim analysis to estimate the variance or the effect size.

Perhaps the most interesting result of this study is that when the sample size per arm is larger than 200, SSR generates no major advantages over the typical design because the typical design itself already offers adequate power. This is in contrast to the results of many finished clinical trials with equally large or even larger sample sizes [3, 4]. The reason for this difference may be that, in our analysis, a moderate effect size was assumed to exist at each measurement and persist from the beginning to the end of each simulated trial (Tables 2 and 3). This difference might indicate that if a moderate clinical meaningful treatment effect indeed persists and can be reflected in the change of ADAS-Cog, it probably won't take a very large sample to detect it. However, in reality, large degree of uncertainty in the effect size or the variance prevents efficient trials with relatively small sample sizes.

Table 2. The average change in ADAS-Cog from baseline by groups and the average difference between the two groups at each visit after the added treatment effect for sample size 50 per arm based on the SL trial

Added effect size	Groups	m1	m3	m6	m9	m12	m15	m18	m21	m24
.25	Placebo	-0.68	0.35	1.8	3.34	5.36	7.05	8.62	9.47	11.29
	Treatment	-1.35	-0.5	0.95	2.00	3.94	5.23	6.63	7.28	8.9
	Difference	0.67	0.86	0.85	1.33	1.42	1.82	1.99	2.19	2.38
.15	Placebo	-0.49	0.56	1.85	3.69	5.77	7.08	8.69	9.62	11.52
	Treatment	-1.1	-0.58	0.88	2.54	4.47	6.14	7.52	7.81	9.9
	Difference	0.60	1.15	0.96	1.15	1.3	0.95	1.175	1.815	1.63

Table 3. The average change in ADAS-Cog from baseline by groups and the average difference between the two groups at each visit after the added treatment effect for sample size 50 per arm based on the HC trial

Added effect size	Group	m3	m6	m9	m12	m15	m18
.25	Placebo	1.28	1.44	3.03	4.25	5.54	6.34
	Treatment	-0.18	0.4	1.1	2.72	4.4	4.47
	Difference	1.46	1.03	1.93	1.53	1.14	1.87
.15	Placebo	1.21	1.41	2.28	3.17	4.79	5.44
	Treatment	0.63	1.04	1.66	2.92	4.7	5.28
	Difference	0.58	0.38	0.62	0.25	0.08	0.17

Although our analysis demonstrates the effectiveness of SSR for relatively small initial sample sizes, there are some limitations that must be considered. First, the gain in power after SSR depends on the initial sample sizes. Though we have recommended SSR for trials with initial sample sizes less than 200 per arm, the optimum pre-trial sample size was not determined. Second, the possible impact of the recruitment rate on the time of SSR has not been investigated. Very fast recruitment rates mean that at the interim analysis, most or even all of the subjects have been enrolled, and it might not be necessary to conduct SSR given a relatively larger initial sample size, e.g. larger than 200 per arm. However, considering failures in completed clinical trials in AD with large sample sizes, SSR can still be used to determine whether to stop larger trials early for futility, or whether to increase the number of longitudinal measurements instead of the number of recruits [19]. Third, unique features of longitudinal trials might be incorporated in SSR in the future. For example, when the recruitment period is shorter than the trial duration, the interim analysis may not contain any complete data. In addition, as the variances of the outcome increase over time, the estimate of the variance at the interim analysis may underestimate the variance of the later time points. Research to address these questions is in progress. Fourth, the flexibility to recruit additional subjects and the gain in power after SSR introduces added complexity of logistics, masking, telegraphy of results, and statistical analysis.

CONCLUSIONS

The SSR adaptive design can be effective for AD and MCI trials with small to medium initial sample sizes. It can not only lead to significant gains in power, but also avoid the exposure of a large number of patients to ineffective treatment by starting the trial with a relatively small initial sample size and stopping the trial early for futility. Considering the need to identify effective treatments, the continuous increase in sample size for AD trials, and the difficulty in estimating pre-trial treatment effects, the SSR adaptive design can be a superior alternative to the typical randomized, placebo-controlled, parallel-group design.

REFERENCES

[1] Thies W, Bleiler L. 2013 Alzheimer's disease facts and figures. Alzheimer's & dementia: the journal of the Alzheimer's Association. 2013;9:208-45.

[2] Aisen P, Andrieu S, Sampaio C, Carrillo M, Khachaturian Z, Dubois B, et al. Report of the task force on designing clinical trials in early (predementia) AD. Neurology. 2011;76:280-6.

[3] Schneider LS, Sano M. Current Alzheimer's disease clinical trials: methods and placebo outcomes. Alzheimer's & dementia. 2009;5:388-97.

[4] Knopman DS. Clinical trial design issues in mild to moderate Alzheimer disease. Cognitive and behavioral neurology: official journal of the Society for Behavioral and Cognitive Neurology. 2008;21:197.

[5] Cummings J, Gould H, Zhong K. Advances in designs for Alzheimer's disease clinical trials. American journal of neurodegenerative disease. 2012;1:205.

[6] Chow S-C, Chang M. Adaptive design methods in clinical trials-a review. Orphanet J Rare Dis. 2008;3.

[7] Kennedy RE, Cutter GR, Schneider LS. Effect of< i> APOE</i> genotype status on targeted clinical trials outcomes and efficiency in dementia and mild cognitive impairment resulting from Alzheimer's disease. Alzheimer's & Dementia. 2013.

[8] DA W. Wechsler memory scale-revised. San Antonio: Psychological Corporation. 1987.

[9] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in medicine. 2006;25:4279-92.

[10] Schneider LS, Kennedy RE, Cutter GR. Requiring an amyloid-β< sub> 1-42</sub> biomarker for prodromal Alzheimer's disease or mild cognitive impairment does not lead to more efficient clinical trials. Alzheimer's & Dementia. 2010;6:367-77.

[11] Petersen RC, Thomas RG, Grundman M, Bennett D, Doody R, Ferris S, et al. Vitamin E and donepezil for the treatment of mild cognitive impairment. New England Journal of Medicine. 2005;352:2379-88.

[12] Doody R, Ferris S, Salloway S, Sun Y, Goldman R, Watkins W, et al. Donepezil treatment of patients with MCI A 48-week randomized, placebo-controlled trial. Neurology. 2009;72:1555-61.

[13] Lawrence Gould A, Shih WJ. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. Communications in Statistics-Theory and Methods. 1992;21:2833-53.

[14] Chang M. Adaptive design theory and implementation using SAS and R: CRC Press; 2007.

[15] Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. Statistics in Medicine. 1990;9:65-72.

[16] Hamer R, Simpson P. Last observation carried forward versus mixed models in the analysis of psychiatric clinical trials. American Journal of Psychiatry. 2009;166:639-41.

[17] Proschan MA. Sample size re- estimation in clinical trials. Biometrical Journal. 2009;51:348-57.

[18] Kairalla JA, Coffey CS, Thomann MA, Muller KE. Adaptive trial designs: a review of barriers and opportunities. Trials. 2012;13:145-.

[19] Shih WJ, Gould AL. Re- evaluating design specifications of longitudinal clinical trials without unblinding when the key response is rate of change. Statistics in Medicine. 1995;14:2239-48.

EFFECT OF SAMPLE SIZE RE-ESTIMATION IN ADAPTIVE DESIGN CLINICAL TRIALS FOR ALZHEIMER'S DISEASE WHEN THE KEY RESPONSE IS THE RATE OF CHANGE

GUOQIAO WANG, RICHARD E. KENNEDY, GARY R. CUTTER, LON S. SCHNEIDER

In preparation for submission

Format adapted for dissertation

1 INTRODUCTION

The number of individuals with AD continues to grow worldwide with the aging of the population [1]. Although a handful of modestly effective symptomatic treatments have been developed using the typical randomized clinical trial (RCT) design, clinical trials to identify effective disease-modifying treatments to slow the progression of AD have been uniformly negative [2-4]. There are several potential causes of these negative trials, including the lack of efficacy in the treatments, insensitivity of the primary outcome to treatment changes, and low power due to the inaccurate pre-trial estimates of the treatment effect. Therefore, clinical trial designs which allow interim analyses and resultant modification of the ongoing trial to increase or adjust power, such as adaptive designs, are recommended [5]. One such adaptive design is the sample size re-estimation (SSR) adaptive design, which allows sample size adjustment based on the comparison between the interim treatment effect (or the interim variance) to the pre-trial treatment effect (or variance) [6].

For longitudinal data, SSR can be conducted based on a single measurement of the primary outcome. For example, the variance of ADAS-Cog at 6 months can be estimated at the interim analysis and then compared to the pre-trial variance, and the ratio of these two will determine the necessity of sample size adjustment (Figure 1). This method can be problematic for at least two reasons: 1) the estimated variance at the interim analysis likely underestimates the variance of the primary outcome at the end of the study since it has been observed that the variance of the longitudinal outcome increases over time [3]; 2) it does not take advantage of the other available measurements at the interim analysis.

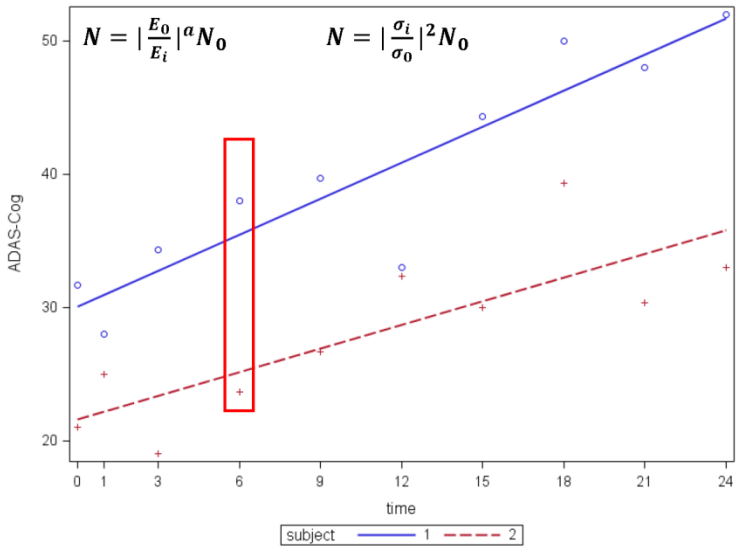

Figure 1. SSR based on a single measurement at 6 months of the primary outcome in a longitudinal study

Alternatively, SSR can also be conducted based on the rate of change of the longitudinal measurements. For example, the variance of the rate of change in ADAS-Cog scores can be estimated at the interim analysis using all the available measurements, and then compared to the pre-trial estimate so that the decision as to whether or not adjust the sample can be made (Figure 2). This method uses all the available measurements at the interim analysis, and its accuracy is not affected by the increasing variance of the primary outcome provided that the model used is correct.

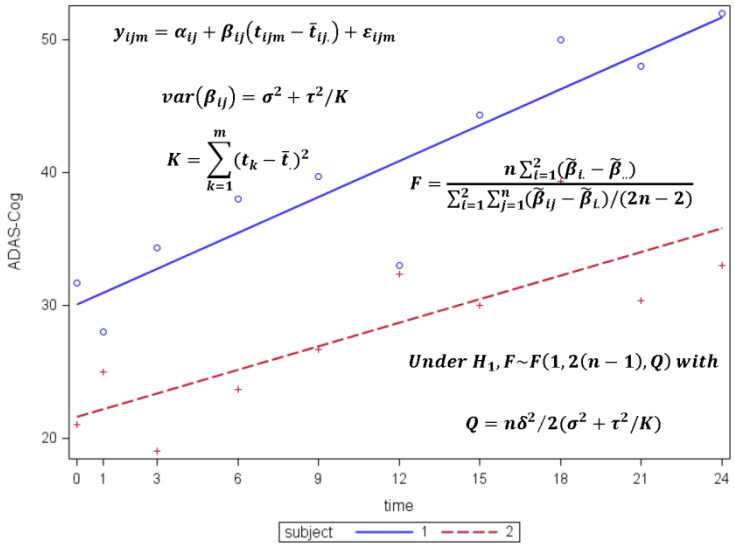

Figure 2. SSR based on the rate of change of the primary outcome in a longitudinal study

The application of SSR based on a single measurement in AD showed that it can increase the power effectively; however, the gain in power varied remarkably due to the poor estimates of the interim effect size and/or the interim variance since only a single measurement was used. This paper examined the advantages of using all the available measurements at the interim analysis in performing SSR based on the rate of change in longitudinal data, which is expected to yield estimates with more accuracy and less variation, using simulations derived from real patient data.

2 METHODS

2.1 Study Overview

Participants for the simulations were drawn from two clinical trials: the vitamins B (HC) using its duration of 18 months, and the selegiline/vitamin E (SL) with duration of 24 months, as well as the pooled data of HC and SL. Clinical assessments were done at 3-month intervals with the exception of 1st-month assessments for SL trial (Table 1). These two trials were chosen from the meta-database because of their longer duration and

more frequent and regular measurements. Patients with missing measurements of 3 or more in the last 5 measurements were excluded in order to avoid heavily underestimated rates of change due to the use of the last observation carry forward (LOCF) imputation method. Since the results are based on simulations, this restriction was not seen as very important despite the potential for informative censoring. After the exclusion, 136 patients from the SL trial (136/341 =40%) and 335 patients from the HC trial (335/459=73%) were selected for simulation.

Table 1. The assessment schedule and the estimates of the pre-trial between-subject and within-subject variances based on the chosen AD clinical trials

Trials	Time of clinical assessments (months)	ES=.15 (within/between)	ES=.25 (within/between)
		SS=100	SS=100
ADNI	0, 6, 12, 24	12.6(1.1)/14.6(3.0)	13.1(1.1)/15.5(2.9)
DHA	0, 6, 12, 18	12.6(1.0)/18.7(3.4)	13.3(1.1)/18.7(3.3)
ES	0, 2, 6, 12, 15	8.6(0.7)/23.4(3.3)	8.9(0.7)/23.9(3.2)
HC	0, 3, 6, 9, 12,15, 18	14.3(1.2)/16.0(2.8)	15.0(1.2)/16.0(2.8)
LL	0, 3, 6, 12, 18, 20	14.4(0.8)/17.0(3.3)	15.1(1.0)/17.6(3.4)
PR	0, 1, 2, 7, 12, 17	8.2(0.7)/17.4(2.1)	8.6(0.7)/17.9(2.1)
SL	0, 1, 3, 6, 9, 12,15, 18, 21, 24	10.9(0.6)/10.5(1.4)	11.4(0.7)/10.6(1.3)
	Mean of the means	Mean: *11.6/16.8*	Mean: *12.2/17.1*

Abbreviations: SS, sample size; ES, effect size; within, the within-subject variance which is the variability of ADAS-cog measurements over time; between, the between-subject variance which is the variability of the patient-specific slopes

The separate and then combined use of data from the two trials with same measurement spacing provided the opportunity to compare the effect of SSR over different trials and durations, and to evaluate the reliability of the pre-trial variances estimated from all of the available clinical trials in our meta-database. The primary outcome chosen for simulation was the ADAS-cog, which evaluates memory, reasoning, orientation, praxis, language, and word finding difficulty, and is scored from 0 to 70

errors [7]. The baseline characteristics of all the clinical trials in the meta-database were shown in Table 2.

Table 2. The baseline characteristics of all the AD clinical trials in the meta-database

Study	N	Age	Education (years)	*Gender (M/F)(%)		*Race (Non White/White)		*Marital Status (Y/N)(%)		**Baseline ADAS-Cog
HC	459	76(8)	14(3)	180 (44.0)	229 (56.0)	104 (22.7)	355 (77.3)	293 (63.8)	166 (36.2)	22.7 (8.8)
SL	341	73(8)	12(3)	119 (34.9)	222 (65.1)	40 (11.7)	301 (88.3)	250 (73.3)	91 (26.7)	30.7 (9.6)

* The percentages of different categories are significantly differently between the studies ($p<.0001$) based on the chi-square test.
**The means of the baseline ADAS-Cog scores are significantly different among the studies even after adjustment for age and education ($p<.0001$) based on the general linear model.

2.2 Estimate of the Pre-trial Variances

For each clinical trial in the meta-database, 100 clinical trials of sample size 100 were simulated, and the within-subject and between-subject variances were estimated and averaged accordingly. The mean of all the means of those clinical trials were considered as the pre-trial within-subject and between-subject variances in the simulation study (Table 1).

2.3 Simulation Principles and Parameters

Simulations were conducted under a detailed protocol [8], similar to our previously published approach [9, 10], to reflect clinical trials for an experimental drug for AD or MCI with one treatment group and one placebo group, and parameters for the distribution of ADAS-Cog selected to be consistent with previously published trials and ADNI [11, 12]. Parameters used to simulate the clinical trials are shown in Table 3.

Table 3. Parameters used in the simulation

Parameters	Scenarios
Data source (duration)*	HC (18), SL (24) and the pooled (18)
Primary outcome	ADAS-Cog
Trial duration**	15/(6, 9)/18, 18/(6, 9, 12)/24
Initial sample size per arm	50, 100, 200
Effect size	0.15, 0.25
Random error in placebo groups	$N(0,1)$
Treatment effect	χ^2
Allocation ratio	1:1
Time of SSR	6 months, 9 months, and 12 months
Pre-trial variances	Average of 7 trials (Table 1)
Interim variance estimate	The pooled-sample-variance with adjustment

*The pooled data of HC and SL are used to increase variability in the outcome over time
**15/(6, 9)/18 means that the HC trial is truncated so that trials of 15 months are simulated, SSR at 6 months and 9 months are conducted, and if warranted, the duration simulated trials can be extended to 18 months.

2.4 Duration of the Simulated Trials and the Time of SSR

In order to allow for an extension in time over the initial duration of the trial after SSR, trials with initial duration 15 months were simulated based on HC trial and the pooled of HC and SL trials. A single SSR was conducted at 6 months and 9 months, and the initial duration was then extended to 18 months if warranted. Trials of initial duration of 18 months were simulated based on SL trial. A single SSR was done at 6 months, 9 months, or 12 months, the initial duration was extended to 24 month if warranted.

2.5 The Placebo Group and the Treatment Group

For each scenario, a separate set of patients was constructed by randomly choosing from the meta-database with replacement, i.e., patients from the dataset could be present in the simulated groups more than once in the same or different treatment groups, but were perturbed with random components, thus lessening the correlations due

to sampling with replacement. The placebo group outcome was the score for the subject at the specified time point in the meta-database, with random error added to minimize ties in the outcome and thus making even the same patient if selected again slightly different. For each subject in the treatment group, effect sizes of 0.15 and 0.25 were used to compute expected treatment results representing treatment effects of small to medium size. The individual treatment effect was randomly generated from a χ^2 distribution with a mean equal to the expected treatment effect (effect size times pooled standard deviation) to allow for a more realistic distribution of declines over time, where a few patients may fail or worsen more markedly than would be predicted by a normal distribution. As higher scores on the ADAS-cog reflect poorer performance, the individual treatment effect was shifted by subtracting two times the expected treatment effect, then adding the resultant to the patient's score at the specified time point in the database. For example, if a is the ADAS-Cog score at a given time point in the database, then $a + \chi_z^2 - 2 * z$, is the corresponding score in the simulated treatment group, where $z = effect\ size * sd$ and sd is the sample standard deviation of the change in ADAS-Cog from baseline. Assume $a = 24$, effect size is 0.25, sd is 8, and the randomly generated treatment effect from the χ_z^2 is 3, then the ADAS-Cog score used in the simulation would be 24 + 3 – 2*0.25*8= 23. While a patient may be reused in the analysis, the actual value used would be modified by this randomly selected amount, hence making it slightly different. The resultant mean differences at each time point between the treatment group and the placebo group after the added treatment effect were shown in table 4 and table 5. The difference between the two groups continued to increase over time, although they are quite variable.

Table 4. The average change in ADAS-Cog from baseline by treatment groups and the average difference between the two groups at each visit after the added treatment effect for sample size 50 per arm based on the SL trial

Added effect size	Groups	*m1	m3	m6	m9	m12	m15	m18	m21	m24
0.25	Placebo	-0.68	0.35	1.8	3.34	5.36	7.05	8.62	9.47	11.29
	Treatment	-1.35	-0.5	0.95	2.00	3.94	5.23	6.63	7.28	8.9
	Difference	0.67	0.86	0.85	1.33	1.42	1.82	1.99	2.19	2.38
0.15	Placebo	-0.49	0.56	1.85	3.69	5.77	7.08	8.69	9.62	11.52
	Treatment	-1.1	-0.58	0.88	2.54	4.47	6.14	7.52	7.81	9.9
	Difference	0.60	1.15	0.96	1.15	1.3	0.95	1.175	1.815	1.63

*m1 represents that the measurement was taken after the first month, and so on

Table 5. The average change in ADAS-Cog from baseline by treatment groups and the average difference between the two groups at each visit after the added treatment effect for sample size 50 per arm based on the HC trial

Added effect size	Group	m3	m6	m9	m12	m15	m18
0.25	Placebo	1.28	1.44	3.03	4.25	5.54	6.34
	Treatment	-0.18	0.4	1.1	2.72	4.4	4.47
	Difference	1.46	1.03	1.93	1.53	1.14	1.87
0.15	Placebo	1.21	1.41	2.28	3.17	4.79	5.44
	Treatment	0.63	1.04	1.66	2.92	4.7	5.28
	Difference	0.58	0.38	0.62	0.25	0.08	0.17

2.6 SSR Method

Clinical trials for AD are generally longitudinal studies in which each patient is followed over a period of time and is repeatedly measured multiple times with even or uneven spacing, leading to a series of measurements in chronological order. For these longitudinal data, the rate of change has been used as a key response or outcome variable [13-15]. The rate of change can be obtained through the following model which was recommended by Aisen [2] and was reproduced based on a paper by Shih [13]. Let y_{ijk}

be the k^{th} measurement for the j^{th} patient in the i^{th} group, $i = 1,2; j = 1, \dots n$, assuming equal allocation in both groups; $k = 1, \dots m$, assuming the same number of measurements for each patient. The outcome measurements y_{ijk} at time t_{ijk} can be related to a patient-specific intercept α_{ij} and a patient-specific slope β_{ij} through the following linear regression model:

$$y_{ijk} = \alpha_{ij} + \beta_{ij}\left(t_{ijk} - \bar{t}_{ij.}\right) + \varepsilon_{ijk},$$

where, $\bar{t}_{ij.} = \sum_{k=1}^{m} \frac{t_{ijk}}{m}$. The error term ε_{ijk} is assumed to be independently and identically distributed (i.i.d) as $N(0, \tau^2)$. In order to facilitate comparison of slopes of the two treatment groups, the patient-specific slope is expressed as the sum of the fixed treatment effect β_i of group i and a random patient effect ϵ_{ij},

$$\beta_{ij} = \beta_i + \epsilon_{ij},$$

where, the random effect term ϵ_{ij} is i.i.d as $N(0, \sigma^2)$. Here τ^2 and σ^2 are referred to as the within-subject measurement error variance and the between-subject variance, respectively.

The null and alternative hypotheses to test the treatment effect are:

$$H_0: \beta_1 = \beta_2, \qquad H_1: \beta_1 \neq \beta_2.$$

The statistic for testing H_0 based on the least-square estimate $\tilde{\beta}_{ij}$ of β_{ij} is the F statistic:

$$F = \frac{\sum_{i=1}^{2} n(\tilde{\beta}_{i.} - \tilde{\beta}_{..})^2}{\sum_{i=1}^{2}\sum_{j=1}^{n}(\tilde{\beta}_{ij} - \tilde{\beta}_{i.})^2/(2n-2)},$$

where $\tilde{\beta}_{i.} = \frac{\sum_j \tilde{\beta}_{ij}}{n}$, and $\tilde{\beta}_{..} = \frac{\beta_{1.}+\beta_{2.}}{2}$

Under H_1, F follows a non-central F distribution with $(1, 2n-2)$ degrees of freedom and non-centrality parameter

$$Q = \frac{n\delta^2}{2\left(\sigma^2 + \frac{\tau^2}{K}\right)} = \frac{n\delta^2}{2var\left(\tilde{\beta}_{ij}\right)},$$

where $K = \sum_{k=1}^{m}(t_{ijk} - \bar{t}_{ij.})^2$, and $\delta = \beta_1 - \beta_2$ refers to the minimal clinically meaningful (important) difference. The power of a non-central F distribution is a monotonically increasing function of Q. To preserve the power of the test, we keep the non-centrality parameter unchanged. Let τ_0^2 and σ_0^2 denote the pre-trial estimate of τ^2 and σ^2, respectively. Suppose that at the interim analysis with n_0 patients per arm, who completed $m_0(\leq m)$ of the m measurements, the estimates of τ^2 and σ^2 are τ_i^2 and σ_i^2. Then in order to preserve the power, we require:

$$Q = \frac{n\delta^2}{2\left(\sigma_0^2 + \frac{\tau_0^2}{K}\right)} = \frac{\tilde{n}\delta^2}{2\left(\sigma_i^2 + \frac{\tau_i^2}{\tilde{K}}\right)}.$$

This implies that we may need to adjust either or both n and K. If the sample size n is to be changed, and K, the number of measurements is to be unchanged, we increase the sample size and retain the duration of the trial, then

$$\tilde{n} = \frac{n(K\,\sigma_i^2 + \tau_i^2)}{K\sigma_0^2 + \tau_0^2}. \qquad (1)$$

If n is to be unchanged, and K is to be changed, meaning to retain the same sample size but increase the duration of the trial, then

$$\widetilde{K} = \frac{\tau_i^2}{\left\{(\sigma_0^2 - \sigma_i^2) + \frac{\tau_0^2}{K}\right\}}. \qquad (2)$$

2.7 Estimates of the Within-subject and Between-subject Variances at the Interim Analysis

Let $N = 2n$ denote the total numbers of patients in the two groups. The estimates of patient-specific intercepts and slopes using the least-square method can be shown to be:

$$\tilde{\alpha}_{ij} = m^{-1} \sum_{k=1}^{m} y_{ijk} = \bar{y}_{ij.},$$

$$\tilde{\beta}_{ij} = K^{-1} \sum_{k=1}^{m} \left(y_{ijk} - \bar{y}_{ij.}\right)\left(t_{ijk} - \bar{t}_{ij.}\right).$$

Then the within-subject measurement variance is estimated by

$$\tilde{\tau}_i^2 = \frac{\sum_{i=1}^{2} \sum_{j=1}^{n} \{y_{ijk} - \tilde{\alpha}_{ij} - \tilde{\beta}_{ij}(t_{ijk} - \bar{t}_{ij.})\}^2 a}{N(m-2)}.$$

The between-subject variance was estimated using a method similar to Lefante's [16] assuming that the two groups have equal variances. First, we estimate the grand mean slope,

$$\tilde{\beta}_{..} = N^{-1} \sum_{i=1}^{2} \sum_{j=1}^{n} \tilde{\beta}_{ij}.$$

Then the between-subject variance and the within-subject variance are related by equating the measures of variability to their expectation,

$$E\left[(N-1)^{-1} \sum_{i=1}^{2} \sum_{j=1}^{n} K(\tilde{\beta}_{ij} - \tilde{\beta}_{..})^2\right] = \tau^2 + \sigma^2/K.$$

That leads to the estimate of

$$\tilde{\sigma}_i^2 = (N-1)^{-1} \sum_{i=1}^{2} \sum_{j=1}^{n} (\tilde{\beta}_{ij} - \tilde{\beta}_{..})^2 - \frac{\tilde{\tau}_i^2}{K}.$$

This estimate of the between-subject variance is biased [17], and overestimates the true variance and thus needs to be adjusted by $\frac{\delta^2}{4} = (\beta_1 - \beta_2)^2$ leading to the unbiased estimate of σ^2 to be $\tilde{\sigma}_i^2 = (N-1)^{-1} \sum_{i=1}^{2} \sum_{j=1}^{n} (\tilde{\beta}_{ij} - \tilde{\mu})^2 - \frac{\tilde{\tau}_i^2}{K} - \frac{\delta^2}{4}$. Once the interim variances are obtained, using formulae (1) and (2), either n and/or K can be re-estimated.

2.8 Estimate of the Drug Effect $\delta = \beta_1 - \beta_2$

No estimate of the difference in mean slopes δ has been provided by the results of any recent AD clinical trials, however, a 40% reduction from the placebo group has been recommended as the minimal clinically meaningful drug effect [18]. Assuming a 40% reduction in the mean slope (β_1) of the placebo group, then the mean slope of the treatment group is $\beta_2 = 0.6\beta_1$. When both groups have the same number of patients, the overall mean slope can be calculated as

$$\beta_{..} = \frac{\beta_1 + \beta_2}{2} = 0.8\beta_1,$$

which implies

$$\delta = 0.4\beta_1 = 0.5\beta_{..}.$$

Thus the estimate of δ is

$$\tilde{\delta} = 0.5 * \frac{\sum_{i=1}^{2} \sum_{j=1}^{n} \tilde{\beta}_{ij}}{N}.$$

3 RESULTS

There exists significant variation among the different AD clinical trials in baseline characteristics such as gender, marital status, race, and baseline ADAS-Cog score. Clinical trial SL had the most severe patients on entry and this may explain the heavy losses to follow-up allowing us to utilize only 40% of the data, which if true would underestimate the true slopes. HC had the least severe patients in terms of baseline ADAS-Cog score (Table 2). The within-subject variance increased slightly, while the between-subject variance decreased remarkably, when the number of measurements used to estimate the slope increased (Table 6). However, the estimates of the mean slopes were stable over the increase of 3 or less measurements used in the estimation. The assumption of equal variances including the equal within-subject variance, the equal between-subject variance, and the equal total variance is satisfied based on the estimates of the pooled data; and this assumption is independent of the measurement spacing and the total number of measurements (Table 7).

Table 6. The stability in the estimates of the within-subject and between-subject variances over the numbers of measurements used

Trials	Time of measurements (months)	(Within/Between) Effect Size=.25
		SS=100
HC	0, 3,6,9,12	14.0/17.2
	0, 3,6,9,12,15	14.8/13.0
	0, 3,6,9,12,15,18	14.9/16.4
SL	0, 1,3,6,9,12	9.5/19.2
	0, 1,3,6,9,12,15,18	10.2/15.0
	0, 1,3,6,9,12,15,18,21,24	11.4/10.3
The pool of HC and SL	0, 3,6,9,12	12.9/18.5
	0, 3,6,9,12,15	13.7/14.0
	0, 3,6,9,12,15,18	13.7/15.9

Table 7. The estimates of the between-subject variances, the within-subject variances, and the total variances of the patient-specific slopes for trials in the meta-database with sample sizes 100 per arm and effect size 0.25

Trials	Time of measurement	Variances(between/within/total)	
		Placebo	Treatment
ES	0, 2, 6, 12, 15	23.7/9.7/31.3	24.2/8.2/32.3
HC	0-18 by 3	16.9/13.8/24.8	16.6/16.1/25.8
LL	0, 3, 6, 12, 18, 20	18.5/13.7/24.6	18.6/16.4/25.7
PR	0, 1, 2, 7, 12, 17	18.8/7.8/24.0	18.5/9.5/24.1
SL	0-24 by 3	12.6/10.3/14.9	12.3/12.5/15.1

When trials were simulated based on HC data and the pooling of both HC and SL data with an initial duration of 15 months, the use of SSR at 12 months with an extension of the duration to 18 months if warranted, the gain in power was greater for adjustments to the sample size than for adjustments in the number of measurements. When only the single study SL was used with an initial duration of 18 months, the use of SSR at 12 months with an extension of the duration to 24 months if warranted, the gain in power for the former was less than for the latter (Figure 3). When an increase in the sample size was required, the number of extra samples needed does not depend on the initial trials used for the simulation as long as the time of SSR remained the same (Figure 4).

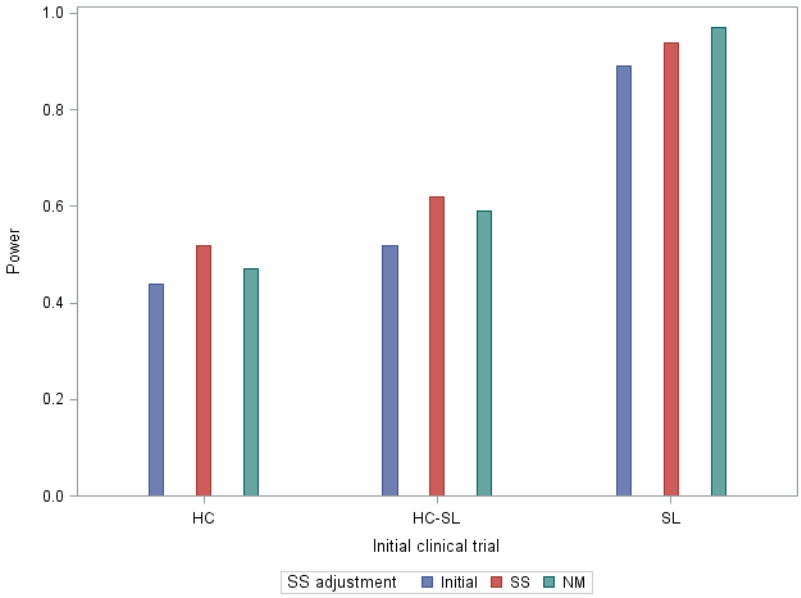

Figure 3. Power comparison before and after sample size adjustment. Initial duration: 15 months, Extension of duration: 18 months, SSR at: 12 months, Interim SSR sample size: 50 per arm, Sample size per arm: 50, Effect size: 0.25, pre-trial variances: 10 and 10 for SL, 11 and 16(between) for HC and HC-SL, NM: number of measurements

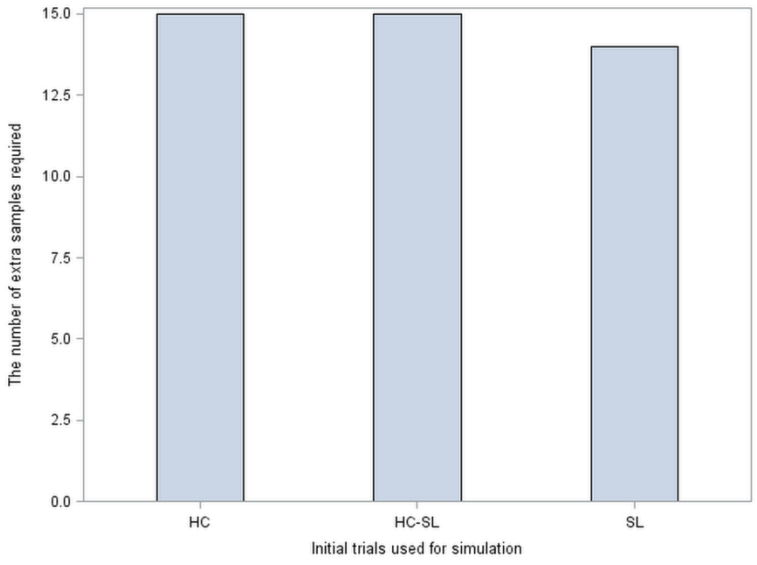

Figure 4. The increase in sample size after SSR at 12 months for trials simulated based on different initial trials, for HC or HC-SL trials: initial duration of 15 months and an extension to 18 month; for SL only trials: initial duration of 18 months and an extension to 24 month

Varying the time of SSR did not significantly affect the gain in power (Figure 5), however, it did result in large differences in the frequency of different types of

adjustment (Figure 6). For trials simulated based on the SL study, SSR at a later time was more likely to lead to adjustments in both the sample size and/or the number of measurements than SSR at an earlier time. When an increase in sample size occurred, the former on average required less extra samples. Similar results were obtained when trials were simulated based on the HC data with an initial duration of 15 months, SSR at 9 months and 12 months, and pre-trial within- and between-subject variances 11 and 16, except that SSR at 9 months actually resulted into more frequent increases in sample size than at 12 months.

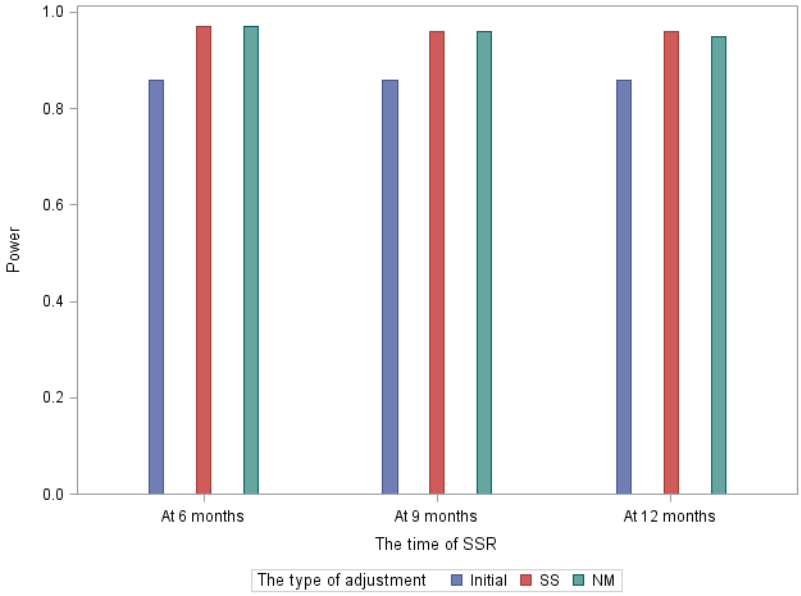

Figure 5. The gain in power after SSR at different time based on SL. Initial duration: 18 months, Extension of duration: 24 months, Interim SSR sample size: 50, SS per arm: 50, Effect size: 0.25, pre-trial variances: 10 and 10

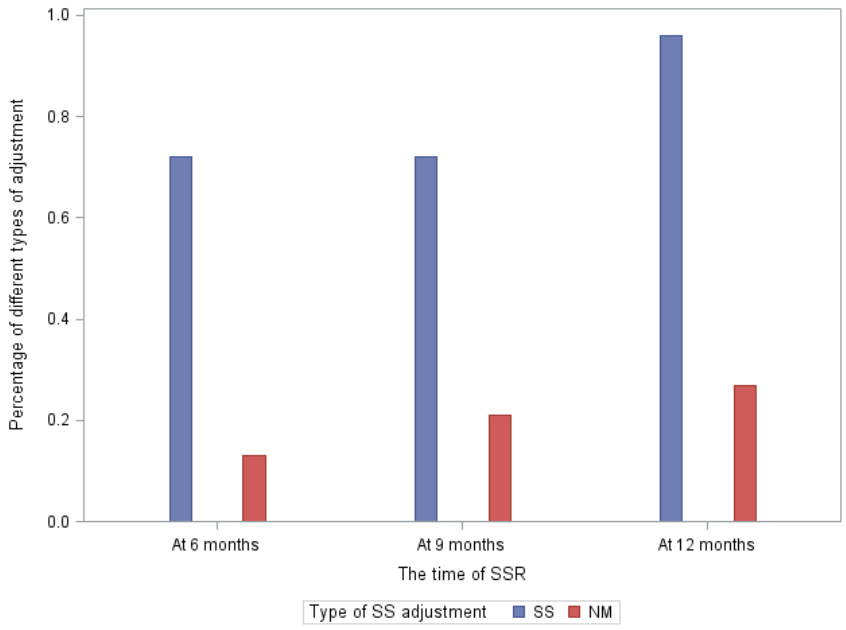

Figure 6. The percentage of different types of adjustment by the time of SSR based on SL. Initial duration: 18 months, Extension of duration: 24 months, Interim SSR sample size: 50, SS per arm: 50, Effect size: 0.25, pre-trial variances: 10 and 10

4 DISCUSSIONS

The application of the SSR adaptive design in AD has evolved as an option to insulate against poor or uninformed planning and may become necessary due to the dominant portion of negative trials with large sample sizes and long durations. If SSR is to be done, this paper extends our previous work to the longitudinal trials when the patient-specific slopes are the key response. The two-stage random effect model was used to allow the estimation of the within-subject and the between-subject variances of the slopes [19]. Both variances were blinded estimated at the interim analysis sparing the Type I error, but allowing for effective increases in sample size or duration.

Based on our study, the SSR based on the rate of change can be effective. It not only increases the power, but also helps to determine the type of sample adjustment. Our

simulation results show that although the SSR can lead to either increase in the sample size or in the number of measurements, the frequency of adjustments falls to increase the sample size is rather higher than that to increase the number of measurements provided the timing of SSR is the same. This is a useful result because the logistics of extending the duration of a trial in terms of adding measurements is far more complicated than adding sample size. This result can be explained by the larger between-subject variances in the pooled trials compared to the within-subject variances and would likely hold in general. Given the same time of SSR, the adjustment in sample size generally leads to slightly more gains in power than that in the number of measurements when the latter only requires an increase of 2 or fewer measurements which would translate to 6 months of trial time even with these minimal increases.

The time of SSR significantly affects the frequency of different types of sample adjustments, particularly the adjustment in the number of measurements. It is interesting that the later the SSR, the more likely the sample adjustment. This result is due to the increase of both the between-subject and the within-subject variances over the increased number of measurements used in the estimation. It also explained why the SSR at a later time requires more extra samples if the sample size to be increased. Therefore, the time of SSR is crucial. The simulation results show that given the number of measurements, the estimates of the slopes and their variances become reasonably stable when the sample size is over 50; on the other hand, given the sample size, the estimates become reasonably stable over 5 or more measurements, which is equivalent to a 12 months trial with one measurement every 3 months. These results were also observed by Shih, et al. [13].

The results obtained in our simulation are very similar regardless of the original data used for the simulation, meaning that not only the SSR method does behave consistently across trials, but also it is safe to use the data of previous trials given that trials employed the same measurement schedule. In addition, the results are also consistent over different treatment effect sizes; particularly, in our simulation, small (.15) to moderate (.25) effect sizes were used.

Despite the effectiveness of SSR for different types of trials, there are some limitations in this study. First, the outcomes of the SSR depends on the pre-trial estimates of the corresponding variances, and how to use the pooled data to get these estimates when the data are not all equally spaced, needs further investigation. Second, the simulation results are based on only two trials with the same measurement schedule (one measurement every 3 months), and they might not be applicable to trials with different measurement schedule. Third, the LOCF imputation method may underestimate the progression rate if missing values are consecutive and are in the end of the study; in addition, missing data usually decrease the power of tests when the key response is the rate of change in the longitudinal data. However, the magnitude of this negative impact has not been investigated [20]. Fourth, although the method to add a treatment effect in this study resulted in a group difference in ADAS-Cog comparable to the existing results, our method also generated a group difference in the early and middle stage of the trial [21]. This consistent group difference in ADAS-Cog scores may or may not reflect what would happen in a real trial. Thus, these results need to be verified through a real trial.

In this study, the SSR is based on the variances of the slopes where the slopes are directly estimated for each patient; an alternative is to conduct the SSR based on the

variance of the interaction coefficient between the treatment and time using generalized estimating equation (GEE) method discussed by Jung [22].

REFERENCES

[1] Thies W, Bleiler L. 2013 Alzheimer's disease facts and figures. Alzheimer's & dementia: the journal of the Alzheimer's Association. 2013;9:208-45.

[2] Aisen P, Andrieu S, Sampaio C, Carrillo M, Khachaturian Z, Dubois B, et al. Report of the task force on designing clinical trials in early (predementia) AD. Neurology. 2011;76:280-6.

[3] Schneider LS, Sano M. Current Alzheimer's disease clinical trials: methods and placebo outcomes. Alzheimer's & dementia. 2009;5:388-97.

[4] Knopman DS. Clinical trial design issues in mild to moderate Alzheimer disease. Cognitive and behavioral neurology: official journal of the Society for Behavioral and Cognitive Neurology. 2008;21:197.

[5] Cummings J, Gould H, Zhong K. Advances in designs for Alzheimer’s disease clinical trials. American journal of neurodegenerative disease. 2012;1:205.

[6] Chow S-C, Chang M. Adaptive design methods in clinical trials-a review. Orphanet J Rare Dis. 2008;3.

[7] DA W. Wechsler memory scale-revised. San Antonio: Psychological Corporation. 1987.

[8] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in medicine. 2006;25:4279-92.

[9] Kennedy RE, Cutter GR, Schneider LS. Effect of< i> APOE</i> genotype status on targeted clinical trials outcomes and efficiency in dementia and mild cognitive impairment resulting from Alzheimer's disease. Alzheimer's & Dementia. 2013.

[10] Schneider LS, Kennedy RE, Cutter GR. Requiring an amyloid-β< sub> 1-42</sub> biomarker for prodromal Alzheimer's disease or mild cognitive impairment does not lead to more efficient clinical trials. Alzheimer's & Dementia. 2010;6:367-77.

[11] Petersen RC, Thomas RG, Grundman M, Bennett D, Doody R, Ferris S, et al. Vitamin E and donepezil for the treatment of mild cognitive impairment. New England Journal of Medicine. 2005;352:2379-88.

[12] Doody R, Ferris S, Salloway S, Sun Y, Goldman R, Watkins W, et al. Donepezil treatment of patients with MCI A 48-week randomized, placebo-controlled trial. Neurology. 2009;72:1555-61.

[13] Shih WJ, Gould AL. Re- evaluating design specifications of longitudinal clinical trials without unblinding when the key response is rate of change. Statistics in Medicine. 1995;14:2239-48.

[14] Dawson JD, Lagakos SW. Analyzing laboratory marker changes in AIDS clinical trials. JAIDS Journal of Acquired Immune Deficiency Syndromes. 1991;4:667-76.

[15] Love RR, Mazess RB, Barden HS, Epstein S, Newcomb PA, Jordan VC, et al. Effects of tamoxifen on bone mineral density in postmenopausal women with breast cancer. New England Journal of Medicine. 1992;326:852-6.

[16] Lefante JJ. The power to detect differences in average rates of change in longitudinal studies. Statistics in medicine. 1990;9:437-46.

[17] Govindarajulu Z. Robustness of sample size re- estimation procedure in clinical trials (arbitrary populations). Statistics in medicine. 2003;22:1819-28.

[18] Aisen P, Andrieu S, Sampaio C, Carrillo M, Khachaturian Z, Dubois Bu, et al. Report of the task force on designing clinical trials in early (predementia) AD. Neurology. 2011;76:280-6.

[19] Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. Longitudinal data analysis: CRC Press; 2008.

[20] Overall JE, Shobaki G, Shivakumar C, Steele J. Adjusting sample size for anticipated dropouts in clinical trials. Psychopharmacology bulletin. 1997;34:25-33.

[21] Grill JD, Di L, Lu PH, Lee C, Ringman J, Apostolova LG, et al. Estimating sample sizes for predementia Alzheimer's trials based on the Alzheimer's Disease Neuroimaging Initiative. Neurobiology of aging. 2013;34:62-72.

[22] Jung SH, Ahn C. Sample size estimation for GEE method for comparing slopes in repeated measurements data. Statistics in medicine. 2003;22:1305-15.

DESIGN PARAMETERS AND EFFECT OF THE DELAYED-START DESIGN FOR ALZHEIMER'S DISEASE

GUOQIAO WANG, RICHARD E. KENNEDY, GARY R. CUTTER, LON S. SCHNEIDER

In preparation for submission
Format adapted for dissertation

1 INTRODUCTION

The number of individuals with AD continues to grow worldwide with the aging of the population [1]. Although a handful of modestly effective symptomatic treatments have been developed using the typical randomized clinical trial (RCT) design, clinical trials to identify effective disease-modifying treatments to slow the progression of AD have been uniformly negative [2-4]. Without any effective treatments to slow down the progression or delay the onset of AD, as many as 7.1 million people in the United States are estimated to live with AD [5]. Disease-modifying treatments should not only ameliorate the symptoms of AD, but also be able to affect the underlying pathology of the disease [6]. However, the disease-modifying effect may not be distinguished from the symptomatic effect using the typical parallel-group design [7]. In order to facilitate the detection of disease-modifying treatments, a number of novel clinical designs have been proposed as the alternative to the standard RCT. One of them is the delayed-start (DS) design (also referred to as the randomized-start design). In the DS design, patients are randomly assigned to placebo or treatment for a pre-specified period of time and then those (or a randomized portion of those) in the placebo group are given the treatment. If patients who are on the treatment from the beginning of the study have similar outcomes to those who received the treatment later, the treatment effect, if any, is considered symptomatic, but not disease-modifying [6, 7].

The DS design has been used in at least one study for Parkinson's disease. Despite the inconclusiveness of the Parkinson's disease study to declare the disease-modifying effect of the tested treatment; it has attracted extensive interest among AD researchers, leading to various proposals to exploit its application in AD [8-12]. Depending on whether or not all the patients in the placebo group receive the treatment at the later stage of the study, the DS design yields two main patterns (Figure 1).

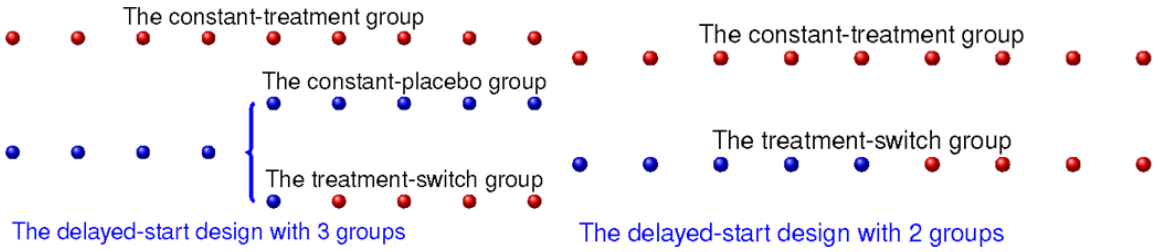

Figure 1. The two main ramifications of the DS design

The one on the left starts with two groups: a treatment group and a placebo group. During the first period of the study, when the goal is to demonstrate the symptomatic effect of the drug, patients are randomly assigned to either receive the drug or the placebo. In the second period of the study, in order to maintain the blinding, a second randomization is performed with the initial placebo group, so that a proportion of the patients would receive the drug and the other would remain on placebo throughout the study [10]. Thus, at the end of the study, there are, in fact, three groups: the constant-treatment group, the constant-placebo group, and the treatment-switch group. The three groups consist of four subsets of patients: those in the constant-treatment group (henceforth referred as the treatment group), those in the constant-placebo group (henceforth referred as the placebo group), those on placebo in the treatment-switch group (henceforth referred as the before-treatment-switch (BTS) group), and those on

treatment in the treatment-switch group (henceforth referred as the after-treatment-switch (ATS) group). The one on the right also starts with the treatment group and the placebo group. However, after the first period of study, all the patients who are initially on placebo would receive the active treatment [8, 9, 11]. The latter design eventually has only two groups (the constant-treatment group and the treatment-switch group) including three subgroups of patients: those in the constant-treatment group, those on placebo in the treatment-switch group, and those on treatment in the treatment-switch group. The two-group design might be preferred in that all the patients on placebo eventually receive the treatment under test. However, because of the lack of effective disease-modifying treatments, putting patients on placebo is still a common practice in the ongoing trials for AD [13] and should yield smaller sample sizes than the active comparative trials. In addition, a placebo group can avoid the risk of bias caused by unblinding.

However, due to the complexity in determining the crucial design parameters such as the sample size allocation ratio in different treatment groups for the three- group DS design, the optimal time of treatment switch, the length of the BTS period, the test statistic, and the power for given sample sizes; the DS design has not been successfully applied in any AD clinical trial to detect disease-modifying treatments [7].

In this study, we extend the investigation of the three- group DS design which was first studied by Xiong [10]. We first provide results as to the sample size allocation ratio among the 3 different groups, the time of treatment switch, the correlation between the primary outcome in the BTS group and the ATS group, and the optimal weight to be used in the final test statistic. We then simulate trials of different durations to compare

the power of the DS design with the typical parallel-group design based on simulated data and on real patient data.

2 METHOD

2.1 Study Overview

This study aimed to improve and propose the design parameters of the DS design and to use simulations based on simulated data and on real patient data to investigate the behavior of the DS design in AD. Data from a meta-database consisting of 5 completed clinical trials (ES, HC, LL, PR, and SL) were used. The primary outcome was the ADAS-cog, which evaluates memory, reasoning, orientation, praxis, language, and word finding difficulty, and is scored from 0 to 70 errors [14]. The spacing of the clinical assessments varied over different clinical trials.

2.2 The Statistical Model

Clinical trials for AD are generally longitudinal studies in which each patient is followed over a long period of time and is repeatedly measured multiple times usually (but not always) with even visit spacing, leading to a series of measurements in chronological order. For these longitudinal data, the rate of change has been suggested as a key response variable [15-17]. Assume that a linear model is appropriate to describe the longitudinal data in both the treatment group and the placebo group, and then the slope over time can be used to measure the rate of change. In addition, assume that the start of the delayed treatment only delays the treatment effect, thus affecting the slope of the longitudinal data. For each group, the rate of change can be evaluated through a two-stage random effects model [18]. The following statistical model was based on papers by

Xiong [10], Lefante [19], and Shih [15]. Let y_{ijk} be the k^{th} measurement for the j^{th} patient in the i^{th} group, $i = t, p, b,$ and a with t representing the treatment group, p the placebo group, b the BTS group, and a the ATS group; $j = 1, \dots, n_i$, meaning that each treatment group may include different number of patients; $k = 1, \dots, m_i$, meaning that patients in different groups may have different numbers of measurements. The outcome measurements y_{ijk} at time t_{ijk} can be related to a patient-specific intercept α_{ij} and a patient-specific slope β_{ij} through the following generalized linear regression model:

$$y_{ijk} = \alpha_{ij} + \beta_{ij}\left(t_{ijk} - \bar{t}_{ij.}\right) + \varepsilon_{ijk},$$

where, $\bar{t}_{ij.} = \left(\sum_{k=1}^{m_i} t_{ijk}\right)/m_i$. The error term ε_{ijk} is independently and identically distributed (i.i.d) with $N(0, \tau^2)$. In order to compare the slopes of the two treatment groups, the patient-specific slope is expressed as the sum of the fixed treatment effect β_i of group i and a random patient effect ϵ_{ij},

$$\beta_{ij} = \beta_i + \epsilon_{ij},$$

where, the random effect term ϵ_{ij} is i.i.d with $N(0, \sigma^2)$. Here τ^2 and σ^2 are referred to as the within-subject measurement error variance and between-subject variance, respectively.

Assume that the treatment switch only affects the rate of change, then the rate of change in the treatment group would be the same as that in the ATS group; and the rate of change in the placebo group would be the same as that in the BTS group. Therefore, the difference between the estimated mean rate of change of the treatment group and that of the placebo group (i.e. $\tilde{\beta}_p - \tilde{\beta}_t$) is an unbiased estimator to the treatment effect ($\beta_p - \beta_t$);

Further, so is the estimated difference between that of the BTS group and that of ATS group. The patient-specific slopes estimated using the least-square method can be shown to be:

$$\tilde{\beta}_{ij} = {K_i}^{-1} \sum_{k=1}^{m_i} \left(y_{ijk} - \bar{y}_{ij.}\right)\left(t_{ijk} - \bar{t}_{ij.}\right),$$

where $K_i = \sum_{k=1}^{m_i} (t_{ijk} - \bar{t}_{ij.})^2$, is referred as the length and frequency of follow-up. Let m_0 denote the measurement from which point a randomized portion of patients in the placebo group start the treatment, which is the same for all patients. Let

$$K_b = \sum_{k=1}^{m_0} (t_k - \bar{t}_b)^2,$$

$$\bar{t}_b = \sum_{k=1}^{m_0} \frac{t_k}{m_0},$$

$$K_a = \sum_{k=m_0}^{m} (t_k - \bar{t}_.)^2,$$

$$\bar{t}_a = \sum_{k=m_0}^{m} \frac{t_k}{m - m_0 + 1}.$$

Then the estimates of patient-specific slopes in the BTS group are

$$\tilde{\beta}_{bj} = {K_b}^{-1} \sum_{k=1}^{m_0} \left(y_{bjk} - \bar{y}_{bj.}\right)\left(t_{bjk} - \bar{t}_b\right),$$

and in the ATS group are

$$\tilde{\beta}_{aj} = {K_a}^{-1} \sum_{m_0}^{m} \left(y_{ajk} - \bar{y}_{aj.}\right)\left(t_{ajk} - \bar{t}_a\right).$$

The estimates of the patient-specific slopes are normally distributed with mean $E\left(\tilde{\beta}_{ij}\right) = \beta_i$ and variances $var\left(\tilde{\beta}_{ij}\right) = \sigma_i^2 + \frac{\tau_i^2}{K_i}$. Moreover,

$$\tilde{\beta}_i = \frac{\sum_j \tilde{\beta}_{ij}}{n},$$

$$\tilde{\tau}_i^2 = \frac{\sum_{j=1}^{n}\{y_{ijk} - \tilde{\alpha}_{ij} - \tilde{\beta}_{ij}(t_{ijk} - \bar{t}_{ij.})\}^2}{n(m_i - 2)},$$

$$\tilde{\alpha}_{ij} = {m_i}^{-1}\sum_{k=1}^{m_i} y_{ijk} = \bar{y}_{ij.},$$

$$\tilde{\sigma}_i^2 = (n-1)^{-1}\sum_{j=1}^{n}(\tilde{\beta}_{ij} - \tilde{\beta}_{\mathrm{i}})^2 - \frac{\tilde{\tau}_i^2}{K_i}.$$

Because patients in the BTS and in the ATS group are the same, their patient-specific slopes are correlated and we assume that $(\tilde{\beta}_{bj}, \tilde{\beta}_{aj})$ follows a bivariate normal distribution with mean $(\beta_b,\ \beta_a)$ and covariance matrix

$$\begin{pmatrix} \mathrm{var}(\beta_{bj}) & \rho\ \mathrm{var}(\beta_{bj})\ \mathrm{var}(\beta_{aj}) \\ \rho\ \mathrm{var}(\beta_{bj})\ \mathrm{var}(\beta_{aj}) & \mathrm{var}(\beta_{aj}) \end{pmatrix}$$

Let $\Delta\tilde{\beta}_{b-a} = \tilde{\beta}_b - \tilde{\beta}_a$ denote the estimated treatment effect from the BTS group and the ATS group, then

$$\sigma_\Delta^2 = var(\Delta\tilde{\beta}_{b-a}) = var(\tilde{\beta}_{bj}) + var(\tilde{\beta}_{aj}) - 2cov(\tilde{\beta}_{bj}, \tilde{\beta}_{aj}).$$

Let the null hypothesis be H_0: $\tilde{\beta}_p = \tilde{\beta}_t$, and the alternative be H_a: $\tilde{\beta}_p - \tilde{\beta}_t = d \neq 0$, then taking advantage of the two estimates of the treatment effect, the test statistic is the combination of two unbiased estimators:

$$T_c = c(\tilde{\beta}_p - \tilde{\beta}_t) + (1-c)(\tilde{\beta}_b - \tilde{\beta}_a), \qquad (1)$$

where $0 < c < 1$ is a constant weight. The variance of T_c is

$$\sigma_{T_c}^2 = \frac{c^2 var(\tilde{\beta}_{pj})}{n_p} + \frac{c^2 var(\tilde{\beta}_{tj})}{n_t} + \frac{(1-c)^2\sigma_\Delta^2}{n_{ba}}, \qquad (2)$$

where, n_p represents the sample size of the placebo group, n_t represents the sample size of the treatment group, and n_{ba} represents the sample size of the BTS group and of the ATS group.

2.3 Consideration in Conducting of a DS Trial

With the model specified by (1) and (2) above, we investigated the following design parameters: 1) the sample size allocation ratio between the treatment group, the placebo group and the treatment-switch group; 2) the time of the treatment switch in the treatment-switch group; 3) the optimal weight c in the test statistic; 4) the estimate of the correlation ρ between the slopes of the BTS group and those of the ATS group as well as its impact on the test statistic; and 5) the assumption on the variances of the estimated slopes in the four groups. In addition, we also compared the power of the DS design to the typical design. In the following sections, we will address them one by one.

3 RESULTS

3.1 The Variance Assumptions

The assumption on the variances of the estimated slopes of the four groups is crucial for determining some of the design parameters. Two different assumptions will be considered in this study.

1) The variances of the estimated slopes of the four groups are equal:

$$\sigma_{tp}^2 = \sigma_{tt}^2 = \sigma_{tb}^2 = \sigma_{ta}^2 = \sigma_{total}^2,$$

where, $\sigma_{tp}^2 = \sigma_p^2 + \frac{\tau_p^2}{K_p}$, $\sigma_{tt}^2 = \sigma_t^2 + \frac{\tau_t^2}{K_t}$, $\sigma_{tb}^2 = \sigma_b^2 + \frac{\tau_b^2}{K_b}$, and $\sigma_{ta}^2 = \sigma_a^2 + \frac{\tau_a^2}{K_a}$.

2) The variances of the estimated slopes of the treatment group and the placebo group are equal; however, they are different from those of the BTS group and the ATS group, which is

$$\sigma_{tp}^2 = \sigma_{tt}^2 \neq \sigma_{tb}^2 = \sigma_{ta}^2.$$

The second assumption was proposed based on what was observed from trials in the meta-database (Table 1). The table showed that the variance of the slopes decreases as the number of measurements increase; and when the time of treatment switch is half way through the trial, the variances of the BTS group and the ATS group are closest to each other.

Table 1. The variances for each group by clinical trials with effect size 0.25 and sample sizes 100 per group

Trial	Time of measurement	OBS	Variances(Between/within)				Overall variance			
			Placebo	Treatment	BTS	ATS	Placebo	Treatment	BTS	ATS
ES	0, 2, 6, 12, 15	3/3	23.7/9.7	24.2/8.2	21.1/6.1	26.2/11.6	31.3	32.3	67.9	65.4
HC	0, 3, 6, 9, 12, 15, 18	3/5	16.9/13.8	16.6/16.1	21.7/11.7	21.6/15.9	24.8	25.8	107.2	47.7
HC		4/4	16.9/13.9	16.3/16.1	29.4/11.7	39.4/14.5	24.9	25.5	66.7	85.7
HC		5/3	16.6/13.8	16.4/16.1	18.2/12.3	56.1/13.7	24.5	25.6	37.8	163.9
LL	0, 3, 6, 12, 18, 20	3/4	18.9/13.8	18.5/16.4	23.7/12.5	17.3/16.8	24.9	25.7	111.9	37.9
LL		4/3	18.5/13.7	18.6/16.4	23.4/12.5	13.4/17.6	24.6	25.7	46.0	76.1
PR	0, 1, 2, 7, 12, 17	3/4	18.8/7.8	18.5/9.5	173.2/4.6	17.4/9.9	23.7	24.4	498.1	28.9
PR		4/3	19.2/7.8	18.5/9.5	36.2/5.2	25.6/9.3	24.0	24.3	62.0	52.4
SL	0, 1, 3, 6, 9, 12, 15, 18, 21, 24	3/8	12.5/10.3	12.3/12.5	99.0/5.5	12.1/12.8	14.8	15.1	267.5	17.0
SL		4/7	12.6/10.3	12.4/12.5	26.3/6.5	12.7/12.8	14.9	15.1	70.5	20.0
SL		5/6	12.5/10.3	12.4/12.5	20.9/7.6	14.3/12.3	14.8	15.2	40.9	25.5
SL		6/5	12.6/10.4	12.2/12.5	21.0/7.9	15.6/12.4	14.9	15.0	31.3	35.5
SL		7/4	12.6/10.3	12.3/12.5	18.1/8.3	28.7/11.3	14.9	15.1	24.2	64.4

3.2. The Sample Size Allocation Ratio between the Treatment Group, the Placebo Group, and the Treatment-switch Group

First, we investigated the design parameters under the assumption of $\sigma_{tp}^2 = \sigma_{tt}^2 = \sigma_{tb}^2 = \sigma_{ta}^2 = \sigma_{total}^2$.

(i) The allocation ratio for two independent and normally distributed samples with equal variances

First, for simplicity, start with only two groups: the treatment group and the placebo group. We want to determine the optimal sample size allocation so that the variance of the test statistic will be minimized. Let N denote the total sample size, let n_S and n_L denote the smaller and the larger sample sizes, respectively. Without loss of generality, let $n_S = tn_L$, $0 < t \leq 1$. When assuming equal variances, for a two-tail test, the formula to calculate the sample size for type I error α and type II error β given that the test statistic follows a normal distribution with a difference δ between the means, can be derived based on Figure 2,

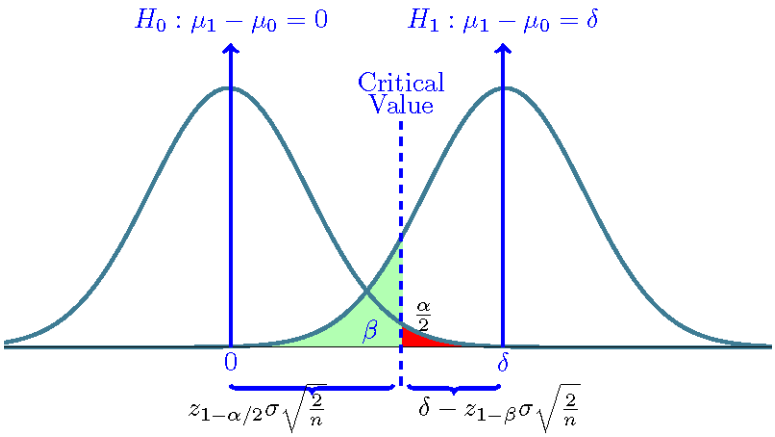

Figure 2. Illustration of sample size calculation for two-tailed test with equal variances

$$Z_{1-\frac{\alpha}{2}}\sigma\sqrt{\frac{1}{n_S}+\frac{1}{n_L}} = \delta - Z_{1-\beta}\sigma\sqrt{\frac{1}{n_S}+\frac{1}{n_L}}.$$

Solving the above equation for n_L, we obtain

$$n_L = \frac{t+1}{t} \times \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 \sigma^2}{\delta^2},$$

thus, the total sample size is

$$N = n_S + n_L = n_L(1+t) = \frac{(t+1)^2}{t} \times \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 \sigma^2}{\delta^2}. \qquad (3)$$

Straightforward mathematical derivation shows that equation (3) reaches its minimum when $t = 1$, indicating that given the total sample size and the pre-determined type I and type II error rates, the equal sample size allocation minimizes the variance of the test statistic, thus leading to maximal power.

(ii) The allocation ratio for two independent and normally distributed samples with unequal variances

When the variances are not equal, without loss of generality, we assume the variance of one group is σ^2 and the variance of the other is $t\sigma^2, 0 < t < 1$. Furthermore, assume the sample means follow normal distributions, then the variance of δ which is the difference between the two sample means, is

$$\frac{\sigma^2}{n_L} + \frac{t\sigma^2}{n_S} = \sigma^2\left(\frac{1}{n_L} + \frac{t}{N - n_L}\right),$$

where, n_L represents the sample size corresponding to the larger variance and n_S represents the sample size corresponding to the smaller one.

In order to maximize the power, the samples need to be allocated in such a way to minimize the variance of δ. Straightforward mathematical calculation shows that the variance is minimized when

$$n_L = \frac{1-\sqrt{t}}{1-t} N. \qquad (4)$$

Which has a limit $n_L = \frac{1}{2}N$, when $t \to 1$. Thus

$$n_S = N - n_L = \frac{\sqrt{t} - t}{1 - t} N. \qquad (5)$$

Thus, the allocation ratio is:

$$n_L : n_S = 1 : \sqrt{t}.$$

The ratio indicates that given the total sample size and the pre-determined type I and type II error rates, in order to maximize power, a larger portion of the samples need to be assigned to the group with the larger variance as might be expected.

(iii) The allocation ratio for two independent and normally distributed samples with unequal variances plus a weight $0 < c < 1$ *such as*

$$T = c\tilde{\mu}_1 + (1 - c)\tilde{\mu}_2,$$

where, $\tilde{\mu}_1$ and $\tilde{\mu}_2$ are the sample means of two normally distributed samples with unequal variance σ^2 and $t\sigma^2, 0 < t \leq 1$. Again, let N denote the total sample size, let n_1 and n_2 denote the sample sizes of the $\tilde{\mu}_1$ group and the $\tilde{\mu}_2$ group, respectively. Then

$$\sigma_T^2 = \frac{c^2\sigma^2}{n_1} + \frac{(1 - c)^2 t\sigma^2}{n_2} = \sigma^2 \left(\frac{c^2}{n_1} + \frac{(1 - c)^2 t}{N - n_1} \right).$$

In order to obtain c and n_1 so that σ_T^2 is minimized, we take partial derivatives with respect to each separately and then solve the following equations simultaneously:

$$c(N - n_1) - tn_1(1 - c) = 0, \qquad (i)$$

$$t(1 - c)^2 n_1^2 - c^2(N - n_1)^2 = 0. \qquad (ii)$$

From (i), we obtain

$$c = \frac{tn_1}{N - n_1 + tn_1}, \qquad (6)$$

and substituting (6) into (ii), yields

$$t - t^2 = 0.$$

On the other hand, solving (ii) first, we obtain

$$n_1 = \frac{N}{1 + \sqrt{t} * \frac{1-c}{c}}, \qquad (7)$$

then substituting (7) into (i), we obtain

$$t - \sqrt{t} = 0.$$

The combination of (6) and (7) indicates that the minimal variance cannot be obtained by solving c and n_1 simultaneously using the partial derivatives. So we try it from another perspective.

Rewrote

$$\sigma_T^2 = \sigma^2 \left(\frac{c^2}{n_1} + \frac{(1-c)^2 t}{N - n_1} \right) = \frac{\sigma^2}{N} \left(\frac{c^2}{{n_1}/{N}} + \frac{(1-c)^2 t}{1 - {n_1}/{N}} \right).$$

In order to reflect the real trial, we restrict $c \in [0.2, 0.8]$ and $\frac{n_1}{N} \in [0.2, 0.8]$, meaning reasonable samples and weights would be put into each group. Then we employ the interior-point algorithm to solve the function iteratively for the minimum given $t \in (0, 1]$ or $t \in [1, \infty)$ [20]. Part of the solutions is listed in table 2. Not surprisingly, the minimums are achieved on the boundaries of c. That is because there are no critical

points in the whole domain, meaning the minimums can only be achieved once boundaries are set. The interior-point algorithm is carried out using Matlab.

Table 2. The combination of c and the sample size allocation ratio to achieve the minimal variance

c	0.2	0.2	0.2	0.2	0.5	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
$\frac{n_c}{N}$	0.36	0.28	0.24	.22	0.5	0.79	0.77	0.76	0.75	0.74	0.73	0.72	0.71	0.705	0.698
t	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3.0

Although, the combinations yield the mathematical minimums, they are not meaningful from the clinical trial standpoint. They basically require the information of c first, and then use c to determine the sample size given the ratio between the variances. However, in a real clinical trial, without knowing the sample size, the trial cannot be carried out, thus resulting in no information of the variances, consequently, c cannot be determined. Therefore, these purely mathematical minimums are not applicable.

Thus, we decide to investigate the allocation from the clinical perspective, meaning we first put certain restrictions on the variances and corresponding sample size allocation ratios. When $t \in (0, 1]$, $\sigma^2 > t\sigma^2$, thus based on our earlier arguments, it is reasonable to allocation more samples (n_1) to the group with larger variance σ^2 in order to minimize the variance, meaning $n_1 \geq n_2$, where, n_2 corresponds to the sample size of the group with the smaller variance $t\sigma^2$. So, $\frac{n_1}{N} \geq \frac{1}{2}$ where, $N = n_1 + n_2$.

First, for any given $c, o < c < 1$, assume that we have

$$c^2\sigma^2 \geq (1-c)^2 t\sigma^2,$$

solving it for c, we obtain $c \geq \frac{1}{1+\frac{1}{\sqrt{t}}}$, and $c \in \left[\frac{1}{1+\frac{1}{\sqrt{t}}}, 1\right)$ for $t \in (0, 1]$.

Applying these results to formula (4),

$$n_1 = n_L = \frac{1 - \frac{1-c}{c}\sqrt{t}}{1 - \frac{(1-c)^2}{c^2}t} N = \frac{c}{c + (1-c)\sqrt{t}} N, \qquad (8)$$

with a limit $\frac{1}{2}N$, when $c \to \frac{1}{1+\frac{1}{\sqrt{t}}}$. That leads to

$$n_2 = n_S = \frac{\frac{1-c}{c}\sqrt{t} - \frac{(1-c)^2}{c^2}t}{1 - \frac{(1-c)^2}{c^2}t} N = \frac{(1-c)\sqrt{t}}{c + (1-c)\sqrt{t}} N. \qquad (9)$$

Substituting both n_S and n_L into $\sigma^2\left(\frac{c^2}{n_L} + \frac{(1-c)^2 t}{N-n_L}\right)$, we obtain

$$\sigma^2\left(\frac{c^2}{n_L} + \frac{(1-c)^2 t}{N-n_L}\right) = \frac{\sigma^2}{N}\left(\frac{c^2}{\frac{1-\frac{1-c}{c}\sqrt{t}}{1-\frac{(1-c)^2}{c^2}t}} + \frac{(1-c)^2 t}{\frac{\frac{1-c}{c}\sqrt{t} - \frac{(1-c)^2}{c^2}t}{1-\frac{(1-c)^2}{c^2}t}}\right)$$

$$= \frac{\sigma^2}{N}\left[\left(c\left(c + (1-c)\sqrt{t}\right)\right) + \left((1-c)\sqrt{t}\left(c + (1-c)\sqrt{t}\right)\right)\right]$$

$$= \frac{\sigma^2}{N}\left(c + (1-c)\sqrt{t}\right)^2 \triangleq f(c).$$

Taking the first derivative of $f(c)$ with respect to c, we obtain

$$\frac{df(c)}{dc} = \frac{2\sigma^2}{N}\left(c + (1-c)\sqrt{t}\right)\left(1 - \sqrt{t}\right) > 0,$$

for any given $c \in \left[\frac{1}{1+\frac{1}{\sqrt{t}}}, 1\right)$ and $t \in (0, 1]$. $f(c)$ is an increasing function of c, thus it achieves the minimum at the lower bound $c = \frac{1}{1+\frac{1}{\sqrt{t}}}$, which leads to $n_L = \frac{1}{2}N$ and $n_S = \frac{1}{2}N$. Therefore, the sample size allocation ratio is 1: 1. Notice that solving $\frac{df(c)}{dc} = \frac{2\sigma^2}{N}\left(c + (1-c)\sqrt{t}\right)\left(1 - \sqrt{t}\right) = 0$ for c gave us $c = \frac{-1}{\frac{1}{\sqrt{t}}-1} < 0$, which is not in the domain $c \in \left[\frac{1}{1+\frac{1}{\sqrt{t}}}, 1\right)$.

On the other hand, if

$$c^2\sigma^2 \leq (1-c)^2 t\sigma^2 \xrightarrow{yields} c \leq (1-c)\sqrt{t},$$

which implies $c \leq \frac{1}{1+\frac{1}{\sqrt{t}}} \leq \frac{1}{2}$, and $\frac{1}{1+\frac{1}{\sqrt{t}}} = \frac{1}{2}$ when $t = 1$. We follow the above arguments and solve for n_1, which is the sample size corresponding to $c^2\sigma^2$, we obtain

$$n_S = n_1 = \frac{\frac{c}{(1-c)\sqrt{t}} - \frac{c^2}{(1-c)^2 t}}{1 - \frac{c^2}{(1-c)^2 t}} N = \frac{c}{(1-c)\sqrt{t} + c} N \leq \frac{1}{2}N,$$

which conflicts with $\frac{n_1}{N} \geq \frac{1}{2}$ when $t \neq 1$. Therefore, this scenario is not appropriate from the clinical perspective because it assigns more samples to the group with the smaller variance.

If $t \in [1, \infty)$, then $\sigma^2 \leq t\sigma^2$, thus it is reasonable to allocate less sample (n_1) to the group with the smaller variance σ^2, meaning $n_1 \leq n_2$, and n_2 corresponds to the sample size of the group with the larger variance $t\sigma^2$. So, $\frac{n_1}{N} \leq \frac{1}{2}$ where, $N = n_1 + n_2$.

First, for any given $c, o < c < 1$, we assume

$$c^2\sigma^2 \leq (1-c)^2 t\sigma^2,$$

Then $c \leq \frac{1}{1+\frac{1}{\sqrt{t}}}$, and $\frac{1}{1+\frac{1}{\sqrt{t}}} \geq \frac{1}{2}$ with $\frac{1}{1+\frac{1}{\sqrt{t}}} = \frac{1}{2}$ when $t = 1$.

Applying formulas (4) and (5) with n_S corresponding to the smaller variance and n_L the larger one, we obtain

$$n_S = n_1 = \frac{\frac{c}{(1-c)\sqrt{t}} - \frac{c^2}{(1-c)^2 t}}{1 - \frac{c^2}{(1-c)^2 t}} N = \frac{c}{(1-c)\sqrt{t} + c} N,$$

$$n_L = n_2 = \frac{1 - \frac{c}{(1-c)\sqrt{t}}}{1 - \frac{c^2}{(1-c)^2 t}} N = \frac{(1-c)\sqrt{t}}{(1-c)\sqrt{t} + c} N.$$

Subsituting both n_S and n_L into $\sigma^2 \left(\frac{c^2}{n_S} + \frac{(1-c)^2 t}{N - n_S}\right)$, we obtain

$$\sigma^2 \left(\frac{c^2}{n_L} + \frac{(1-c)^2 t}{N - n_L}\right) = \frac{\sigma^2}{N} \left(\frac{c^2}{\frac{\frac{c}{(1-c)\sqrt{t}} - \frac{c^2}{(1-c)^2 t}}{1 - \frac{c^2}{(1-c)^2 t}}} + \frac{(1-c)^2 t}{\frac{1 - \frac{c}{(1-c)\sqrt{t}}}{1 - \frac{c^2}{(1-c)^2 t}}} \right)$$

$$= \frac{\sigma^2}{N} \left[\left(c\left(c + (1-c)\sqrt{t}\right) \right) + \left((1-c)\sqrt{t}\left(c + (1-c)\sqrt{t}\right) \right) \right]$$

$$= \frac{\sigma^2}{N}\left(c + (1-c)\sqrt{t}\right)^2 \triangleq f(c).$$

Taking the first derivative of $f(c)$ with respect to c, we obtain

$$\frac{df(c)}{dc} = \frac{2\sigma^2}{N}\left(c + (1-c)\sqrt{t}\right)\left(1 - \sqrt{t}\right) < 0,$$

for any given $c \in \left(0, \frac{1}{1+\frac{1}{\sqrt{t}}}\right]$ and $t \in [1, \infty)$. That means $f(c)$ is a decreasing function of c, thus it achieves the minimum at the upper bound $c = \frac{1}{1+\frac{1}{\sqrt{t}}}$, which corresponds to $n_L = \frac{1}{2}N$ and $n_S = \frac{1}{2}N$. Therefore, the sample size allocation ratio is 1: 1. Notice that solving $\frac{df(c)}{dc} = \frac{2\sigma^2}{N}\left(c + (1-c)\sqrt{t}\right)\left(1 - \sqrt{t}\right) = 0$ for c gave us $c = \frac{-1}{\frac{1}{\sqrt{t}}-1} < 0$, which is not in the domain $c \in \left(0, \frac{1}{1+\frac{1}{\sqrt{t}}}\right]$.

On the other hand, if

$$c^2\sigma^2 \geq (1-c)^2 t\sigma^2 \xrightarrow{yields} c \geq (1-c)\sqrt{t},$$

which means $c \geq \frac{1}{1+\frac{1}{\sqrt{t}}} \geq \frac{1}{2}$, and $\frac{1}{1+\frac{1}{\sqrt{t}}} = \frac{1}{2}$ only when $t = 1$. Applying formulas (4) and (5), solving for n_1, we obtain

$$n_1 = n_L = \frac{1 - \frac{1-c}{c}\sqrt{t}}{1 - \frac{(1-c)^2}{c^2}t}N = \frac{c}{c + (1-c)\sqrt{t}}N \geq \frac{1}{2}N,$$

which conflicts with $\frac{n_1}{N} \leq \frac{1}{2}$ when $t \neq 1$. Therefore, this scenario is not appropriate from the clinical perspective because it assigns fewer samples to the group with the larger variance.

(iv) The optimal sample size allocation ratio and the optimal weight in the test statistic of the DS design $T_c = c(\tilde{\beta}_p - \tilde{\beta}_t) + (1-c)(\tilde{\beta}_b - \tilde{\beta}_a)$.

In the DS design, the first part of the equation, the estimated difference $\tilde{\beta}_p - \tilde{\beta}_t$ can be considered as the estimated mean difference between the treatment group and the placebo group with equal variances, thus, the optimal allocation ratio is 1: 1. Under the assumption of $\sigma_{tp}^2 = \sigma_{tt}^2 = \sigma_{tb}^2 = \sigma_{ta}^2 = \sigma_{total}^2$, $var(\tilde{\beta}_{pj} - \tilde{\beta}_{tj}) = 2\sigma_{total}^2$. Because the slopes in the BTS group and those in the ATS group were calculated based on observations of the same patients, they are correlated. Assume the correlation is ρ, then $var(\tilde{\beta}_{bj} - \tilde{\beta}_{aj}) = 2(1-\rho)\sigma_{total}^2$, thus $t = 1 - \rho$. The test statistic can be considered as a sum of two estimated treatment effects from two samples with unequal variances. The first sample is the estimated treatment effect from the treatment group and the placebo group; and the second is the estimated treatment effect from the BTS group and the ATS group. Assume the sample sizes of the treatment group, the placebo group, and the treatment-switch group, are n_1, n_1, and n_2 respectively; in addition, assume $n_1 + n_2 = N$.

When $\rho < 0$, $t = 1 - \rho < 1$, the allocation ratio is

$$n_1 : n_1 : n_2 = 1 : 1 : 1.$$

Moreover, this allocation ratio also yields an optimal test when comparing the mean slopes from the treatment group and the ATS group under the null hypothesis:

$$H_0: \tilde{\beta}_t - \tilde{\beta}_a = 0,$$

and the alternative:

$$H_a: \tilde{\beta}_t - \tilde{\beta}_a \neq 0.$$

In sum, under the assumption of $\sigma_{tp}^2 = \sigma_{tt}^2 = \sigma_{tb}^2 = \sigma_{ta}^2 = \sigma_{total}^2$, the optimal sample size allocation ratio for the test statistic $T_c = c(\tilde{\beta}_p - \tilde{\beta}_t) + (1-c)(\tilde{\beta}_b - \tilde{\beta}_a)$, is 1: 1: 1.

3.3 The Time of Treatment Switch in the Treatment-switch Group

The time of treatment switch affects the variances in both the BTS group and the ATS group; and it should be chosen to minimize $var(\tilde{\beta}_{bj} - \tilde{\beta}_{aj})$ in order to achieve maximal power given the same sample size. It has been shown that the optimal switch point is the middle one if the measurements are evenly spaced and the total number of measurements is odd; and the optimal switch point is either one of the middle two measurements if the measurements are evenly spaced and the total number of measurements is even [10] (Figure 3).

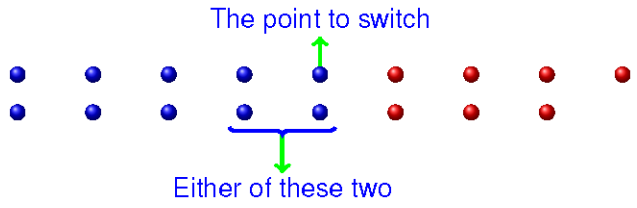

Figure 3. Illustration of the time of treatment switch derived theoretically by Xiong

We estimated the variances of the two groups varying the number of measurements in each group, and the results were presented in Table 1. From Table 1, it is shown that in order to satisfy the equal variance assumption, the optimal switch is the middle measurement when the number of the total number of measurements is odd; and

is the second one of the middle two measurements when the total number of measurements is even regardless the measurement spacing. Therefore, based on the theoretical results and the practical calculation, we recommend that in order to minimize the total variance and meet the equal variance assumption, the optimal switch point should be the middle one when the number of the total number of measurements is odd; and is the second one of the middle two when it is even, regardless whether the measurement spacing is regular or irregular (Figure 4).

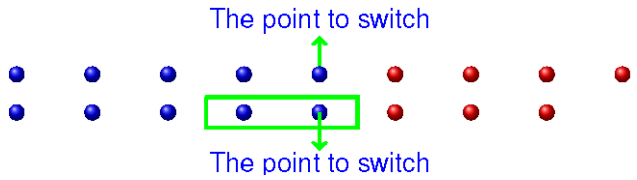

Figure 4. Illustration of the time of treatment switch

3.4 The Optimal Weight c in Equation (1)

Under the assumption of equal variances among the four groups, $t = 1 - \rho$, thus the optimal weight $c = \frac{1}{1+\frac{1}{\sqrt{t}}} = \frac{1}{1+\frac{1}{\sqrt{1-\rho}}}$ with a limit $\frac{1}{2}$ as $\rho \to 0$.

3.5 The Correlation ρ between the Slopes in the BTS Group and Those in the ATS Group

For each clinical trial with 6 or more measurements, we simulated with replacement, 100 replicates of the treatment-switch group with sample sizes 100 from the individual trials in our meta-database. The means and the standard deviations of ρ were estimated. Despite the differences in the duration, number of measurements, and the measurement spacing, ρ is uniformly small with an upper bound 0.20; in addition, more than half of the estimates are negative (Table 3).

Table 3. The estimates of the correlation ρ between the slopes of the BTS group and the ATS group based on different clinical trials

Trial	Time of measurement	#	Positive Mean(SD)	Negative Mean(SD)	*Ratio of +/-	Mean slopes (BTS)	Mean slopes (ATS)
ES	0, 2, 6, 12, 15	3	-0.15(0.08)	0.06(.04)	20/80	4.607(0.963)	1.290(0.794)
HC	0, 3, 6, 9, 12, 15, 18	3	-0.10(.06)	.06(.05)	23/77	3.421(0.983)	3.237(0.652)
		4	-0.20(.11)	.06(.06)	14/86	3.336(0.847)	2.519(0.888)
		5	-0.13(.08)	.08(.05)	37/63	3.724(0.501)	0.744(1.431)
LL	0, 3, 6, 12, 18, 20	3	-0.15(.1)	.09(.08)	25/75	3.906(1.291)	3.770(0.950)
		4	-0.17(.11)	.08(.06)	22/78	4.755(0.984)	2.142(1.210)
PR	0, 1, 2, 7, 12, 17	3	-0.15(.1)	0.13(.1)	37/63	1.151(3.303)	3.050(0.700)
		4	-0.18(.1)	.09(.08)	25/75	3.776(1.057)	2.157(0.937)
SL	0, 1, 3, 6, 9, 12, 15, 18, 21, 24	3	-0.1(.08)	.09(.07)	37/63	3.820(1.292)	5.092(0.634)
		4	-0.1(.07)	0.1(.06)	40/60	4.840(0.909)	4.643(0.646)
		5	-0.11(.08)	.08(.07)	51/49	5.725(0.850)	3.916(0.790)
		6	-0.15(.09)	.07(.06)	23/77	5.999(0.862)	3.145(1.124)
		7	-0.22(.12)	.07(.05)	23/77	6.191(0.622)	1.443(1.274)

*"+" represent the positive correlation, and "-" negative.

3.6 Power Comparison between Three Designs based on Simulated Data

The random intercepts (also referred as patient-specific intercepts) are simulated from a normal distribution:

$$\alpha_{ij} = N(0,1) + baseline\ mean.$$

The random slopes (also referred as patient-specific slopes) are also simulated from a normal distribution:

$$\beta_{ij} = N(0,1) * \sigma + \beta_i.$$

The primary outcomes are calculated based on the simulated intercepts and slopes:

$$y_{ijk} = \alpha_{ij} + \beta_{ij} t_{ijk} + N(0,1) * \tau.$$

For these simulated outcomes, its variance is

$$Var(y_{ijk}) = t^2{}_{ijk}\sigma^2 + \tau^2,$$

and the variance of the corresponding slopes is

$$Var(\beta_{ij}) = \sigma^2 + \frac{\tau^2}{K}.$$

Suppose, the DS trial has a sample size of $3n$, meaning n per group. Then we use "the small typical trial" to refer a randomized, placebo-controlled trial with sample size of n per group; and we use "the large typical trial" to refer a randomized, placebo-controlled trial with sample size of $\frac{3n}{2}$ per group.

First, we compare the power of the DS trial, the small typical trial, and the large typical trial when $\tau^2 = 0$, meaning no within-subject error (Figure 5). For the DS trial, the

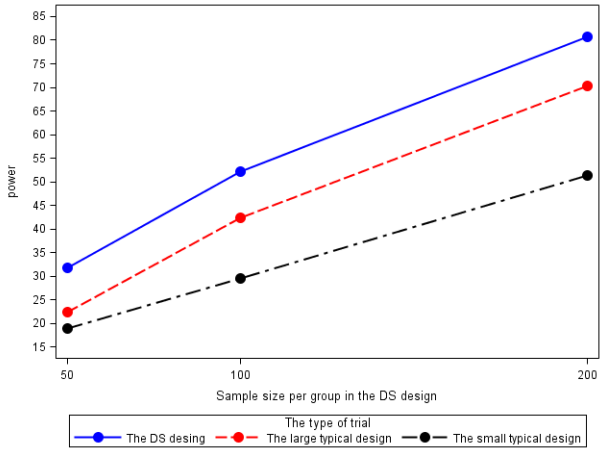

Figure 5. Power comparison between the DS trial (sample size: $n{:}\,n{:}\,n$), the large typical trial (sample size: $\frac{3n}{2}{:}\,\frac{3n}{2}$), and the small typical trial (sample size: $n{:}\,n$) when within-subject error is 0. Given the total sample size, the DS trial has more power than the typical large trial.

test statistic are the ratio of (1) to (2). Formula (1) yields un unbiased estimate of the treatment effect. Under the assumptions of $\tau^2 = 0$, $\rho = 0$ and $c = 0$, formula (2) yields:

$$\sigma^2_{T_c} = c^2 \frac{\sigma^2}{n} * 4 = \frac{\sigma^2}{n}.$$

The test statistic for the large typical design is:

$$Z_c = \frac{\tilde{\beta}_p - \tilde{\beta}_t}{\sigma^2{}_{Typical}},$$

where,

$$\sigma^2{}_{Typical} = \frac{\sigma^2}{\frac{3n}{2}} * 2 = \frac{\sigma^2}{n} * \frac{4}{3}.$$

So, the large typical trial has larger variance than the DS trial, consequently, leading to less power.

In order to evaluate the impact of within-subject error on power of the three types of trials, we simulate longitudinal trials of 24 months duration with measurement spacing 3 months, $\rho = 0$ and $c = 0$. Under these conditions, $K_{24} = 3.75$; the optimal time of treatment switch is at 12 months, and $K_{12} = .625$. It is shown that when the within-subject error increases, as expected, the power of all the three types of trials decreases. The decrease in power is quicker for the DS trial than the typical trials. When $\frac{\tau}{\sigma} \geq 0.8$, the DS trial no longer has power advantage over the large typical design (Figure 6). We further investigate the cutting point where the DS trial starts to have less power than the

large typical design by equating the variance of the test statistic of the DS trial to that of the large typical trial under the aforementioned conditions:

$$\frac{2c^2(\sigma^2+\frac{\tau^2}{K_{24}})}{n}+\frac{2(1-c)^2(\sigma^2+\frac{\tau^2}{K_{12}})}{n}=\frac{2(\sigma^2+\frac{\tau^2}{K_{24}})}{\frac{3n}{2}},$$

solving it for $\frac{\tau}{\sigma}$, we obtain:

$$\frac{\tau}{\sigma}\cong .76.$$

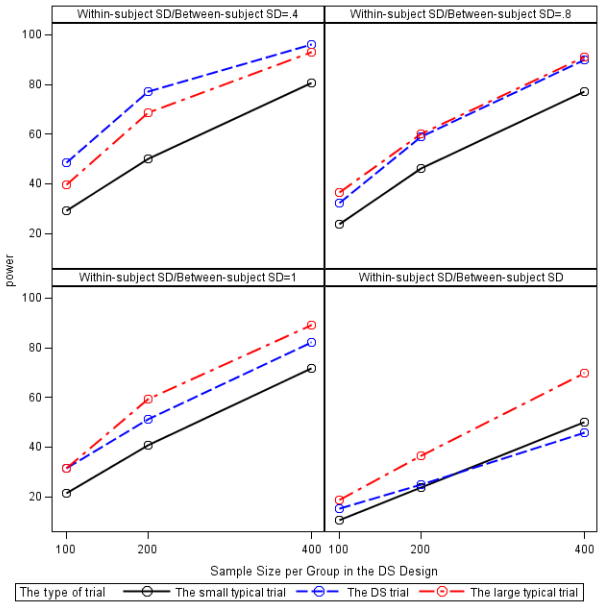

Figure 6. Power comparison between the DS trial (sample size: $n{:}\,n{:}\,n$), the large typical trial (sample size: $\frac{3n}{2}{:}\,\frac{3n}{2}$), and the small typical trial (sample size: $n{:}\,n$) when within-subject error increases.

So far, we have shown that given the pre-trial estimates of the between-subject variance and the within-subject variance, and the trial design parameters such as the duration and the measurement spacing, we can determine whether or not the DS trial would have large power than the large typical design.

3.7 Power Comparison between Three Designs when Assuming Equal Variances based on Real Patient Data

Assume that the goal is to detect the treatment difference using the rate of change as the key response through both the typical parallel-group trials and the DS trials. Then the null and alternative hypotheses for the typical trials are:

$$H_0: \beta_p - \beta_t = 0,$$

$$H_a: \beta_p - \beta_t \neq 0;$$

and the null and alternative hypotheses for the DS trials are:

$$H_0: c\left(\tilde{\beta}_p - \tilde{\beta}_t\right) + (1-c)\left(\tilde{\beta}_b - \tilde{\beta}_a\right) = 0,$$

$$H_a: c\left(\tilde{\beta}_p - \tilde{\beta}_t\right) + (1-c)\left(\tilde{\beta}_b - \tilde{\beta}_a\right) \neq 0.$$

Patients with missing measurements of 3 or more in the last 5 measurements are excluded in order to avoid inaccurate estimates of the rate of change. After the exclusion, 136 of 341 patients from the SL trial and 335 of 459 patients from the HC trial are remained for simulation.

As in the previous section, three types of trials were simulated under a detailed protocol [21], similar to our previously published approach [22] [23], to reflect clinical trials for an experimental drug for AD, and design parameters for the distribution of ADAS-Cog selected to be consistent with previously published trials and ADNI [24, 25]. The key simulation parameters for the three types of trials are presented in table 4. The goal is to compare the power between the three types of trials. Trials simulated based on the SL trial have duration of 24 months and the treatment is switched at 12 months. Trials simulated based on the HC trial have duration of 18 months and the treatment is switched

at 9 months. Trials simulated based on the truncated SL trial only have duration of 18 months and the treatment is switched at 9 months.

Table 4. The key simulation parameters for the 3 different types of trials

Parameters	Values
Assumption of variance	$\sigma_{tp}^2 = \sigma_{tt}^2 = \sigma_{tb}^2 = \sigma_{ta}^2 = \sigma_{total}^2$
ρ	$0, 0.2, 0.4$
c	$c = \frac{1}{1 + \frac{1}{\sqrt{1-\rho}}}$, $c \to 0.5$ $when$ $\rho \to 0$
Allocation ratio	$1:1:1$
Trial designs in comparison	1 treatment-placebo (n_1 per arm) 2 delayed-start (n_1, n_1, n_1) 3 treatment-placebo ($\frac{3n_1}{2}$ per arm)

Trials simulated based on the SL trial with duration of 24 months showed that the power of the DS trials is significantly larger than that of the other two. As expected, when the positive correlation increases, so is the power. However, trials simulated based on the truncated SL trial with duration of 18 month showed the opposite, so are the trials simulated based on the HC trial with the same duration (Figure 7). This conflict might be attributed to the smaller mean slope in the ATS group of trials simulated based on the SL trial (Table 5). The smaller mean slope is more likely caused by the informative dropout, meaning that sicker patients dropped out in the last 6 months and only the healthier remained in the study, thus leading to less progression in ADAS-cog than would be expected. For trials simulated based on the HC trial and the truncated SL trial, the difference in the power between the typical trials and the DS trials increases over sample sizes.

Table 5. The mean slopes and their corresponding variances by initial trials and by trial groups

The initial trial	HC		SL		SLS	
	Slope	Variance	Slope	Variance	Slope	Variance
The treatment group	4.3	24.9	5.9	14.9	6.1	21.7
The placebo group	3.2	25.4	5.2	15.3	5.1	22.2
The BTS group	3.3	66.9	5.8	34.5	4.9	45.5
The ATS group	2.6	85.5	3.8	35.3	4.4	51.0

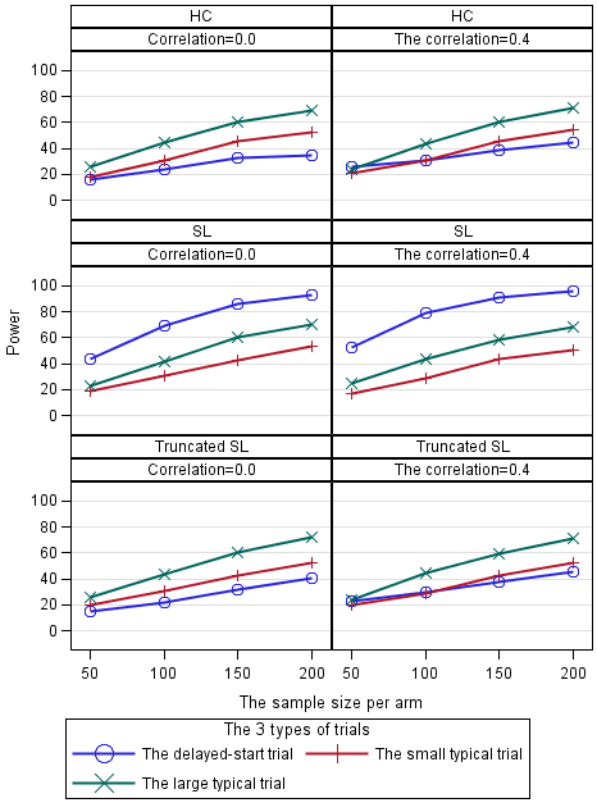

Figure 7. Power comparison between the 3 types of trials by sample sizes and by the original trials used for the simulation

3.7 Under the Assumption of $\sigma_{tp}^2 = \sigma_{tt}^2 \neq \sigma_{tb}^2 = \sigma_{ta}^2$

Assume $\sigma_{tp}^2 = \sigma_{tt}^2 = \sigma_{total}^2$. Based on what is observed from the meta-database, assume $\sigma_{tb}^2 = \sigma_{ta}^2 = s\sigma_{total}^2, s \geq 1$.

The test statistic is the same,

$$T_c = c\left(\tilde{\beta}_p - \tilde{\beta}_t\right) + (1-c)\left(\tilde{\beta}_b - \tilde{\beta}_a\right), \qquad (1)$$

where c $(0 < c < 1)$ is a constant weight. The variance of T_c is

$$\sigma_{T_c}^2 = \frac{c^2 var\left(\tilde{\beta}_{pj}\right)}{n_p} + \frac{c^2 var\left(\tilde{\beta}_{tj}\right)}{n_t} + \frac{(1-c)^2\sigma_\Delta^2}{n_{ba}}. \qquad (2)$$

In the context of the AD data, it can be further simplified by the facts that $var\left(\tilde{\beta}_{pj} - \tilde{\beta}_{tj}\right) = 2\sigma_{total}^2$ and $\sigma_\Delta^2 = var\left(\tilde{\beta}_{bj} - \tilde{\beta}_{aj}\right) = 2(1-\rho)s\sigma_{total}^2$, where ρ is the correlation between $\tilde{\beta}_{bj}$ and $\tilde{\beta}_{aj}$, and is mostly negative based on the meta-database. After simplification, it yields

$$\sigma_{T_c}^2 = \frac{2c^2\sigma_{total}^2}{n_1} + \frac{2(1-c)^2(1-\rho)s\sigma_{total}^2}{n_2} = 2\sigma_{total}^2\left(\frac{c^2}{n_1} + \frac{(1-c)^2(1-\rho)s}{n_2}\right).$$

This matches the scenario which yielded the results (11) and (12), thus by the same arguments, we obtain:

1) The sample size allocation ratio is $1:1:1$;
2) The optimal weight $c = \frac{1}{1+\frac{1}{\sqrt{t}}}$ with $t = (1-\rho)s$;
3) The optimal treatment-switch time is the same as that under the assumption of equal variances;
4) So are the estimate of the correlation;

In order to compare the power of the DS trials with that of large typical trials under the assumption of unequal variances, we simulate trials following the same

protocol used under the assumption of unequal variances. The key simulation parameters for the two types of trials were presented in table 6. Trials simulated based on the HC trial have duration of 18 months and the treatment is switched at 9 months. Trials simulated based on the truncated SL trial have duration of 18 months and the treatment is switched at 9 months.

Table 6. The key simulation parameters for the 2 types of trials

Parameters	Values
Assumption of variances	$\sigma_{total}^2 = \sigma_{tp}^2 = \sigma_{tt}^2 \neq \sigma_{tb}^2 = \sigma_{ta}^2 = s\sigma_{total}^2$
s	$s = 3$ for HC trials, $s=2$ for truncated SL trials
ρ	$0, -0.2, -0.4$
c	$c = \frac{1}{1 + \frac{1}{\sqrt{(1-\rho)s}}}$
Allocation ratio	$1:1:1$
Sample sizes	Large typical trials $(\frac{3n}{2}:\frac{3n}{2})$ Delayed-start trials $(n:n:n)$

The simulation results are the same as those under the equal variance assumption. Again, the DS trials generally had less power than the large typical trials, and the difference between them increased over sample sizes (Figure 8). However, with the right assumption for AD trial, the power of the DS trials increased. And the gain in power also increased over larger sample sizes, meaning that the impact of misspecification of the variance assumption on power increases over the increase of sample sizes (Figure 9).

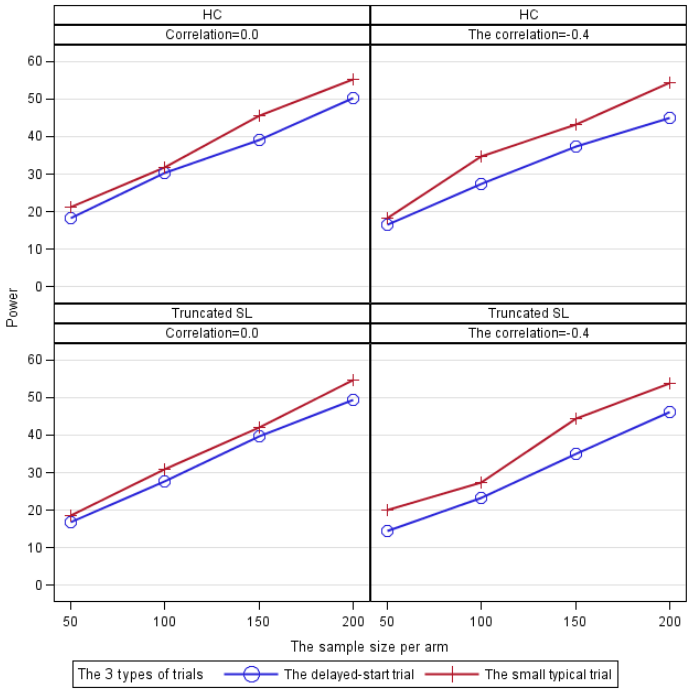

Figure 8. Power comparison between the DS trials and the large typical trials by sample sizes and by the original trials under the unequal variance assumption

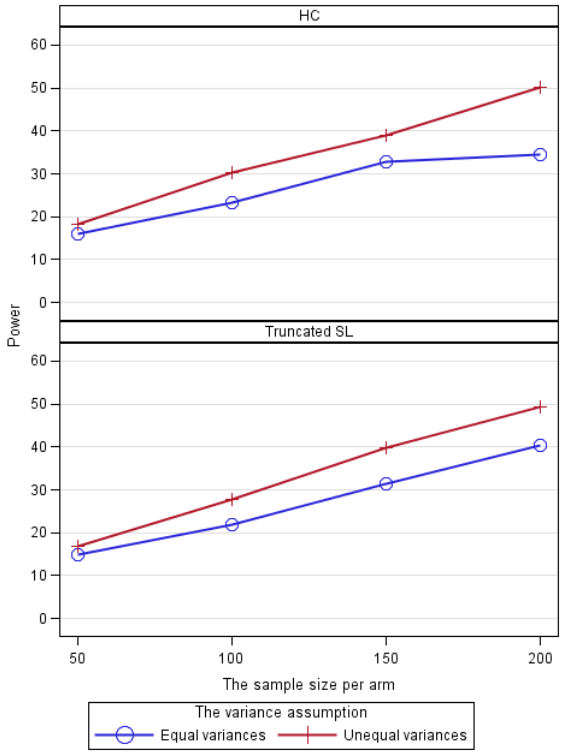

Figure 9. Power of the DS trials under the equal variance assumption and the unequal variance assumption by sample sizes

4 DISCUSSIONS

In this study, we extended the investigation of the DS design with three groups: the constant-treatment group, the constant-placebo group, and the treatment-switch group [10]. First, we obtained the optimal weight c and the optimal sample size allocation ratio through the interior-point algorithm and showed that although the combination of these two yields the mathematical minimum variance of the test statistic, they are not applicable from the clinical trial standpoint because the combination requires knowing c in order to calculate the optimal sample size allocation ratio, which is contradicted to the order of conducting clinical trials in that in a real clinical trial, the sample size allocation ratio has to be determined first to begin the trial. Next, from the clinical trial standpoint, we theoretically proved the optimal sample size allocation ratio and its corresponding optimal weight in the test statistic. We proposed a simple and closed formula for the optimal weight. Our result showed that the optimal sample size allocation ratio is $1:1:1$, which was first obtained by Xiong through simulation [10]. The same allocation ratio was also used in the only DS trial in Parkinson's disease [26], which, however, employed only the constant-treatment group and the treatment-switch group. We also showed that the optimal weight actually varies over the range of correlations, which is different from the conclusion made by Xiong that "the optimum c is only minimally changed when the correlation varies in a wide interval between 0.2 and 0.8" [10].

Based on a meta-database of 7 completed trials in AD, we estimated the absolute correlation to be less than 0.2. More than half of them are actually negative, which have not been considered in the previous study [10]. If the treatment under investigation is

effective, patients' cognition should stop progressing or even get better after they are switched from placebo to treatment. Thus the rate of change in the ADAS-Cog scores will decrease or even change from positive to negative, implying the negative correlations are very likely to happen. The mixture of both negative and positive correlations makes it hard to pre-estimate the correlation in order to calculate the sample size for a future trial. We suggested that data from previously completed trials be used to evaluate the magnitude of the correlation. On the positive side of this mixture, a new type of DS design including only the treatment-switch group with the correlation as the primary outcome may be considered. Under this setting, if the treatment is effective, the rates of change in the ADAS-Cog when patients are on placebo should be negatively correlated with those after patients are switched to the active treatment. So a negative correlation with a large magnitude may be enough to declare the treatment as at least symptomatically effective. A major advantage of this design is that all the patients would receive the treatment, thus it may be better perceived by both patients and trial personnel [11]. However, a downside is the confounding of treatment with the increased decline that occurs over time, potentially making it more difficult to detect treatment effects. Apparently, more research is needed to determine the proper sample size, the test statistic, and the cutting point for a significant negative correlation.

We also compared the power of the DS designs to that of the typical design with only the treatment group and the placebo group based on simulated data. We proved that when the within-subject error is too large, the DS design has no advantages to the typical large design. Moreover, we also proposed the cutting point to decide whether or not the DS design yields more power than the typical design with the same sample size given the

other design parameters such as the duration, the measurement spacing, the estimated optimal weight, and the estimated correlation. This finding conflicts with the conclusion that the DS design generally requires a larger sample size to obtain adequate power compared to the typical design [7, 10, 12]. However, we notice that the DS design with only two groups probably requires larger sample size due to the multiple comparisons, which

Under the assumption that the rate of change is linear and the effect of the treatment switch is only on the rate, it is statistically reasonable to assume that the variances of the four groups are equal. However, what is observed from the meta-data is that the variances of the treatment group and the placebo group are equal, those of the BTS group and the ATS group are equal, and the former is much smaller than the latter. Furthermore, the simulation based on the meta-data showed that the variances are equal as long as the number of measurements in each group is equal; the fewer of the measurements, the larger is the variance. This observation promoted the investigation of the impact of the underlying assumption on the power of the DS design. When assuming the unequal variance, the DS design gains more power, and the gain increases significantly with sample sizes. Considering that the variance assumption not only affects the power, but also determines some of the design parameters, an interim analysis at the point of treatment switch may be necessary.

There are some limitations in this study. First, our study is not able to determine the optimal duration of a DS design. Additionally, the impact of the duration on the power has not been definitely quantified. Second, the simulation is based on only two trials with the same measurement schedule, so whether or not the results are generalizable

may be debatable. Third, whether or not the simulation results are associated with the baseline characteristics such as age, and gender, severity of the dementia has not been investigated. Forth, our simulation results need to be verified in a real trial. Finally, we only investigate the DS design with three arms. Some of the design parameters are not applicable for DS designs with only two arms or even one.

REFERENCES

[1] Thies W, Bleiler L. 2013 Alzheimer's disease facts and figures. Alzheimer's & dementia: the journal of the Alzheimer's Association. 2013;9:208-45.

[2] Aisen P, Andrieu S, Sampaio C, Carrillo M, Khachaturian Z, Dubois B, et al. Report of the task force on designing clinical trials in early (predementia) AD. Neurology. 2011;76:280-6.

[3] Schneider LS, Sano M. Current Alzheimer's disease clinical trials: methods and placebo outcomes. Alzheimer's & dementia. 2009;5:388-97.

[4] Knopman DS. Clinical trial design issues in mild to moderate Alzheimer disease. Cognitive and behavioral neurology: official journal of the Society for Behavioral and Cognitive Neurology. 2008;21:197.

[5] Association zs. 2013 Alzheimer's disease facts and figures. Alzheimer's & dementia: the journal of the Alzheimer's Association. 2013;9:208.

[6] Cummings JL. Defining and labeling disease-modifying treatments for Alzheimer's disease. Alzheimer's & Dementia. 2009;5:406-18.

[7] Mani RB. The evaluation of disease modifying therapies in Alzheimer's disease: a regulatory viewpoint. Statistics in Medicine. 2004;23:305-14.

[8] Olanow CW, Rascol O, Hauser R, Feigin PD, Jankovic J, Lang A, et al. A double-blind, delayed-start trial of rasagiline in Parkinson's disease. New England Journal of Medicine. 2009;361:1268-78.

[9] D'Agostino Sr RB. The delayed-start study design. New England Journal of Medicine. 2009;361:1304-6.

[10] Xiong C, van Belle G, Miller JP, Morris JC. Designing clinical trials to test disease-modifying agents: application to the treatment trials of Alzheimer's disease. Clinical Trials. 2011;8:15-26.

[11] Zhang RY, Leon AC, Chuang-Stein C, Romano SJ. A new proposal for randomized start design to investigate disease-modifying therapies for Alzheimer disease. Clinical Trials. 2011;8:5-14.

[12] Schneider L, Mangialasche F, Andreasen N, Feldman H, Giacobini E, Jones R, et al. Clinical trials and late- stage drug development for Alzheimer's disease: an appraisal from 1984 to 2014. Journal of internal medicine. 2014;275:251-83.

[13] trials.gov C. Clinical trials.gov. 2014.

[14] DA W. Wechsler memory scale-revised. San Antonio: Psychological Corporation. 1987.

[15] Shih WJ, Gould AL. Re- evaluating design specifications of longitudinal clinical trials without unblinding when the key response is rate of change. Statistics in Medicine. 1995;14:2239-48.

[16] Dawson JD, Lagakos SW. Analyzing laboratory marker changes in AIDS clinical trials. JAIDS Journal of Acquired Immune Deficiency Syndromes. 1991;4:667-76.

[17] Love RR, Mazess RB, Barden HS, Epstein S, Newcomb PA, Jordan VC, et al. Effects of tamoxifen on bone mineral density in postmenopausal women with breast cancer. New England Journal of Medicine. 1992;326:852-6.

[18] Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. Longitudinal data analysis: CRC Press; 2008.

[19] Lefante JJ. The power to detect differences in average rates of change in longitudinal studies. Statistics in medicine. 1990;9:437-46.

[20] Thapa GBDMN. Linear programming. 2003.

[21] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in medicine. 2006;25:4279-92.

[22] Kennedy RE, Cutter GR, Schneider LS. Effect of< i> APOE</i> genotype status on targeted clinical trials outcomes and efficiency in dementia and mild cognitive impairment resulting from Alzheimer's disease. Alzheimer's & Dementia. 2013.

[23] Schneider LS, Kennedy RE, Cutter GR. Requiring an amyloid-β< sub> 1-42</sub> biomarker for prodromal Alzheimer's disease or mild cognitive impairment does not lead to more efficient clinical trials. Alzheimer's & Dementia. 2010;6:367-77.

[24] Petersen RC, Thomas RG, Grundman M, Bennett D, Doody R, Ferris S, et al. Vitamin E and donepezil for the treatment of mild cognitive impairment. New England Journal of Medicine. 2005;352:2379-88.

[25] Doody R, Ferris S, Salloway S, Sun Y, Goldman R, Watkins W, et al. Donepezil treatment of patients with MCI A 48-week randomized, placebo-controlled trial. Neurology. 2009;72:1555-61.

[26] Olanow CW, Hauser RA, Jankovic J, Langston W, Lang A, Poewe W, et al. 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. 2008;23:2194-201.

CONCLUSIONS

Summary

Most randomized parallel-group clinical trials to detect symptomatic treatments or disease-modifying treatments have been negative. Novel adaptive designs and the delayed-start (DS) design are perceived to have advantages. How well these novel designs behave in AD clinical trials has never been investigated. Additionally, some key design parameters in the DS design need careful evaluation. Important contributions of this dissertation include evaluating the effect of these novel designs using meta-database of completed AD trials and proposing values for some key design parameters for the DS design.

We presented a thorough comparison between two main designs which allow the use of accumulating data for adaptation of an ongoing trial: the group sequential design and the adaptive design (specifically the SSR adaptive design). We justified that the adaptive design is more fitting, and the GSD should be avoided or used with caution for AD trials.

Different SSR methods were evaluated. SSR using only a single measurement is effective for small or moderate initial sample sizes. However, it generates large variation in both the sample size increases and gains in power. SSR based on the effect size generally results into greater gains in power than SSR based on the variance, however the

former also tends to overshoot the final sample size. SSR at 6 months and SSR at 12 months led to very similar results. But maybe due to LOCF, trials of 24 months did not have more power than those of 18 months; neither did SSR at 12 months than SSR at 6 months. The reason is that 6-month progression in AD cannot overcome the heterogeneity of patients' response to those treatments under investigation. Due to the limitation of the meta-data, the optimal SSR time using only a single measurement was not determined. Considering the increase in variances of the primary outcome over time in a longitudinal study, SSR at 12 months might be more reliable. In summary, when using only a single measurement, SSR based on the variance at 12 months is recommended for trials with small or moderate initial sample sizes.

When SSR was based on the variance of the rate of change in the longitudinal data, it can not only increase the power, but also provide the flexibility between increasing the sample size and increasing the number of measurements. For SSR at the same time, the frequency for the former is much larger than for the latter. Therefore, it is crucial to be realistic about the right length of the trial from the beginning since not only is obtaining more measurements from the same group of patients potentially more difficult than recruiting more patients but also SSR is less likely to lengthen the trial. For this SSR method, the time of SSR still does not impact the gain in power, but it does significantly affect the frequency of sample adjustments including both the sample size and the number of measurements. This speaks volume of the merit of taking advantage of all the information available at the interim analysis. Overall, this method in our opinion is preferable than SSR based on a single measurement, however this conclusion depends on the assumption that a linear model (or mixed effects linear model) is appropriate for

modeling the longitudinal outcome in AD. One main concern about this method is that two variances instead of one need to be determined beforehand, which increases the likelihood of inaccurate estimates.

For the DS design, we not only improved some of the design parameters which were proposed mostly based on simulation, but also extended the variance assumption. A key finding is that violation of the variance assumption could significantly undermine the power of a DS design. We also proposed a cutting point for determining whether or not the DS design has larger power than the typical design given the same sample size. This cutting point is calculated from the pre-trial estimates of the between-subject variance and the within-subject variance, the duration of the trial, the measurement spacing, and the estimate of the correlation between the slopes in the treatment-switch group. This information is also required for planning a typical randomized and placebo-controlled trial. Therefore, our proposal provides the advantage to determine the applicability of the DS design without any extra information. It is widely perceived that the DS design generally needs a larger sample than the typical randomized parallel-group design. The proposed cutting point demonstrates that whether or not the DS design requires a larger sample depends on how many groups the DS design contains. For the DS design with three groups investigated in this dissertation, it is not necessarily true. However, for DS design with only two groups, it is probably true due to the multiple comparisons.

Overall, this dissertation has demonstrated that the novel designs can be effective, and should be employed in future trials. We discovered that the SSR based on the rate of change is less vulnerable to the heterogeneous response of AD patients, and is preferable

if the linear model is indeed the appropriate model. The DS design should be used when the within-subject error is relatively small compared to the between-subject error.

Future Research

SSR based on the variance of the interaction between time and treatment may be a good starting point for future research. For this method, the variance can be estimated using GEE method. This method not only takes advantage of the longitudinal data, but also requires only one pre-trial estimate of the variance. Although it requires assuming a covariance pattern, GEE is robust to misspecification of this pattern.

We started our simulation with a large meta-database for SSR based on the variance or the effect size of a single measurement; however, we ended up only using two trials for SSR based on the variance of the rate of change and for the DS design due to the limitations in trial duration and in the total number of measurements. Thus pooling more trials with equal measurement spacing and long durations to verify the generality of our results might be insightful. Furthermore, with more trials at hand, the impact of baseline characteristics such as age, education, race, and severity of dementia should be examined.

A crucial assumption in SSR based on the variance of the rate of change and in the DS design is that the longitudinal data follows a simple linear trend. So alternatively, nonlinear models could be investigated. However, in order to apply those models, first, it is important to know the turning point of the rate of change in the longitudinal data. So it is helpful to pool more data from trials with long durations to decide where the turning point is and whether or not it is affected by the baseline characteristics.

For the DS design, there are different patterns. We will investigate more of them and provide comparison so that the most appropriate one may be chosen.

Despite all the clinical trials conducted for AD, there is no consensus in the best way to calculate the sample size or power. It will be helpful to have a review paper on this subject.

While this dissertation begins a systematic evaluation of two types of novel designs, as mentioned earlier, there is more to be work on.

GENERAL LIST OF REFERENCES

[1] NIA. Alzhermer's disease-a deeper understanding. 2010.

[2] Thies W, Bleiler L. 2013 Alzheimer's disease facts and figures. Alzheimer's & dementia: the journal of the Alzheimer's Association. 2013;9:208-45.

[3] Petersen RC, Smith GE, Waring SC, Ivnik RJ, Tangalos EG, Kokmen E. Mild cognitive impairment: clinical characterization and outcome. Archives of neurology. 1999;56:303-8.

[4] Cummings JL. Defining and labeling disease-modifying treatments for Alzheimer's disease. Alzheimer's & dementia. 2009;5:406-18.

[5] Knopman DS. Clinical trial design issues in mild to moderate Alzheimer disease. Cognitive and behavioral neurology: official journal of the Society for Behavioral and Cognitive Neurology. 2008;21:197.

[6] Zhang RY, Leon AC, Chuang-Stein C, Romano SJ. A new proposal for randomized start design to investigate disease-modifying therapies for Alzheimer disease. Clinical Trials. 2011;8:5-14.

[7] Medicines in Development For Alzheimer's Disease. America's biopharmaceutical research companies. 2012.

[8] Leber P. Guidelines for the clinical evaluation of antidementia drugs. 1990.

[9] Cummings J, Gould H, Zhong K. Advances in designs for Alzheimer's disease clinical trials. American journal of neurodegenerative disease. 2012;1:205.

[10] Chow S-C, Chang M. Adaptive design methods in clinical trials-a review. Orphanet J Rare Dis. 2008;3.

[11] Mani RB. The evaluation of disease modifying therapies in Alzheimer's disease: a regulatory viewpoint. Statistics in Medicine. 2004;23:305-14.

[12] Cummings JL. Cognitive and behavioral heterogeneity in Alzheimer's disease: seeking the neurobiological basis. Neurobiology of aging. 2000;21:845-61.

[13] Wei L. The adaptive biased coin design for sequential experiments. The Annals of Statistics. 1978:92-100.

[14] Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheiro J. Adaptive designs in clinical drug development—an executive summary of the PhRMA working group. Journal of biopharmaceutical statistics. 2006;16:275-83.

[15] Chang M. Adaptive design theory and implementation using SAS and R: CRC Press; 2007.

[16] Coffey CS, Kairalla JA. Adaptive Clinical Trials. Drugs in R & D. 2008;9:229-42.

[17] Kieser M, Friede T. Re- calculating the sample size in internal pilot study designs with control of the type I error rate. Statistics in medicine. 2000;19:901-11.

[18] Proschan MA, Wittes J. An improved double sampling procedure based on the variance. Biometrics. 2000;56:1183-7.

[19] Govindarajulu Z. Robustness of sample size re- estimation procedure in clinical trials (arbitrary populations). Statistics in medicine. 2003;22:1819-28.

[20] Glimm E, Läuter J. Some notes on blinded sample size re-estimation. arXiv preprint arXiv:13014167. 2013.

[21] Lawrence Gould A, Shih WJ. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. Communications in Statistics-Theory and Methods. 1992;21:2833-53.

[22] Vandemeulebroecke M. Group sequential and adaptive designs–a review of basic concepts and points of discussion. Biometrical Journal. 2008;50:541-57.

[23] Wald A. Sequential analysis. 1947. Wiley, New York; 1947.

[24] Tsiatis AA, Mehta C. On the inefficiency of the adaptive design for monitoring clinical trials. Biometrika. 2003;90:367-78.

[25] Lai TL, Lavori PW, Shih M-C. Adaptive trial designs. Annual review of pharmacology and toxicology. 2012;52:101-10.

[26] Shih WJ. Group sequential, sample size re- estimation and two- stage adaptive designs in clinical trials: a comparison. Statistics in medicine. 2006;25:933-41.

[27] Lawrence Gould A, Shih WJ. On the inappropriateness of an EM algorithm based procedure for blinded sample size re- estimation by T. Friede and M. Kieser, Statistics in Medicine 2002; 21: 165- - 176. Statistics in Medicine. 2005;24:147-54.

[28] Waksman JA. Assessment of the Gould- Shih procedure for sample size re- estimation. Pharmaceutical statistics. 2007;6:53-65.

[29] Fleishman AI. A method for simulating non-normal distributions. Psychometrika. 1978;43:521-32.

[30] Luo H. Generation of Non-normalData: A Study ofFleishman's PowerMethod. 2011.

[31] Kennedy RE, Cutter GR, Schneider LS. Effect of< i> APOE</i> genotype status on targeted clinical trials outcomes and efficiency in dementia and mild cognitive impairment resulting from Alzheimer's disease. Alzheimer's & Dementia. 2013.

APPENDIX A

INSTITUTIONAL REVIEW BOARD APPROVAL

THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Institutional Review Board for Human Use

DATE: November 22, 2013

MEMORANDUM

TO: Guoqiao Wang
Principal Investigator

FROM: Marilyn Doss
Vice Chair
Institutional Review Board for Human Use (IRB)

RE: Request for Determination—Human Subjects Research
IRB Protocol #N130802007 – Application of Longitudinal Data Based Adaptive Design for Alzheimer's Disease and Mild Cognitive Impairment

A member of the Office of the IRB has reviewed your Application for Not Human Subjects Research Designation for above referenced proposal.

The reviewer has determined that this proposal is **not** subject to FDA regulations and is **not** Human Subjects Research. Note that any changes to the project should be resubmitted to the Office of the IRB for determination.