ELSEVIER

Alzheimer's & Dementia: Translational Research & Clinical Interventions 1 (2015) 63-71

Alzheimer's & Dementia

Effect of sample size re-estimation in adaptive clinical trials for Alzheimer's disease and mild cognitive impairment

Guoqiao Wang[a,*], Richard E. Kennedy[b], Gary R. Cutter[a], Lon S. Schneider[c]

[a]*Department of Biostatistics, University of Alabama-Birmingham, Birmingham, AL, USA*
[b]*Department of Medicine, University of Alabama-Birmingham, Birmingham, AL, USA*
[c]*Departments of Psychiatry and Neurology, University of Southern California Keck School of Medicine, Los Angeles, CA, USA*

Abstract

Introduction: The sample size re-estimation (SSR) adaptive design allows interim analyses and resultant modifications of the ongoing trial to preserve or increase power. We investigated the applicability of SSR in Alzheimer's disease (AD) trials using a meta-database of clinical studies.
Methods: Based on six studies, we simulated clinical trials using Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-Cog) as primary outcome. A single SSR based on effect sizes or based on variances was conducted at 6 months and 12 months. Resultant power improvement and sample size adjustments were evaluated.
Results: SSR resulted in highly variable outcomes for both sample size increases and power improvement. The gain in power after SSR varies by initial sample sizes, trial durations, and effect sizes.
Conclusions: SSR adaptive designs can be effective for trials in AD and mild cognitive impairment with small or medium initial sample sizes. However, SSR in larger trials (>200 subjects per arm) generates no major advantages over the typical randomized trials.

Keywords: Alzheimer's disease; Alzheimer's Disease Assessment Scale; Mild cognitive impairment; Sample size re-estimation; Adaptive design

1. Introduction

The number of individuals with Alzheimer's disease (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 have been uniformly negative [2]. 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 pretrial estimates of the treatment effect. Therefore, clinical trial designs allowing interim analyses and the resultant modification of the ongoing trial to increase power have been recommended [3]. 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 pretrial treatment effect (or the pretrial variance) [4].

The typical RCT design starts with a prespecified sample size, and modifications 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 pretrial treatment effect size was overestimated or the pretrial 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 can allow early stopping or an overall decrease in the sample size when the pretrial treatment effect size was underestimated or the pretrial variance was overestimated, leading to a trial with the prespecified power but a smaller sample size,

*Corresponding author. Tel.: +1-205-826-8967; Fax: +1-205-975-2541.
E-mail address: guoqiao@uab.edu

http://dx.doi.org/10.1016/j.trci.2015.03.002

although this is rare. This flexibility not only improves efficacy, but also provides other advantages over the RCT design, such as minimizing the number of patients exposed to inferior treatments, avoiding long-term trials for drugs with limited efficacy, and better using the most recent external or internal information of the ongoing trial. Potential concerns about using SSR include the reliability in estimating the overall treatment effect based on relatively small interim samples (or, for longitudinal trials, the precision in predicting the final treatment effect using only the early measurements), and the trade-off between the gain in power versus the burden to recruit more subjects. The former concern is particularly relevant for AD trials, as heterogeneity in the course of the disease may introduce significant inaccuracies in estimating the final treatment effect based on interim analyses. This study used simulations based on real patient data to investigate the SSR adaptive designs for AD trials.

2. Methods

2.1. Study overview

Of the 19 studies in our meta-database [5], we excluded seven studies that did not collect Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-Cog) data, four trials with duration less than 18 months, one trial enrolling only normal subjects, and one trial enrolling only moderate AD, yielding six studies that were used for the simulation (Supplementary Table 1). 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 [6]. Clinical assessments were obtained at 6-month intervals over the first 2 years.

2.2. Simulation methods

Simulations were conducted under a detailed protocol [7], similar to our previously published approach [5,8], 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 were selected to be consistent with previously published trials and Alzheimer's Disease Neuroimaging Initiative (ADNI) ADNI [9,10].

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, subjects were randomly selected from the meta-database with replacement, i.e., subjects from the data set could be present 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 a 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 than would be predicted by a normal distribution. This method introduced some extreme measurements but not to the extent of violating the homoscedasticity assumption of the analysis models. 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 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, 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. In this example, if 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 in the simulation would be 23. With this added treatment effect, the mean difference in ADAS-Cog between treatment arms and its standard deviation increased over time (Supplementary Table 2).

2.3. Time points used for SSR

Patients' enrolment times vary in a typical trial, leading to different numbers of available measurements for each patient at the interim analysis. In this study, for trials with given initial sample sizes, "SSR at 12 months" means that all the patients had enrolled and had been measured for up to at least 12 months. We truncated the follow-up at the specified time point so our results would not depend on the recruitment rate.

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 for both treatment arms. The "variance only method" assumes that the pretrial estimate of the mean difference between treatment arms is accurate, and only the variance is uncertain and needs re-estimation. At the interim analysis, the variance of ADAS-Cog was estimated and compared with the pretrial estimate, and then the sample size was adjusted based on the following equation [11]:

$$N=\frac{\widehat{\sigma}_i^2}{\widehat{\sigma}_0^2}N_0,$$

where N is the re-estimated sample size, N_0 is the initial sample size, and $\widehat{\sigma}_i^2$ and $\widehat{\sigma}_0^2$ are the interim and the estimated

pretrial variances, respectively. In our analysis, $\widehat{\sigma}_i^2$ was estimated using pooled data (to mimic blinding to treatment in a clinical trial) as $\widehat{\sigma}_i^2=(N_i-1)/(N_i-2)(S^2-\Delta^2/4)$, where N_i is the total sample size at the interim analysis, S^2 is the pooled sample variance, and Δ is the pretrial estimate of the treatment effect [12]. This method does not inflate type I error, thus no adjustment to the α level is required.

The "effect size method" assumed that the pretrial estimates of the mean difference between treatment arms and its variance are uncertain. At the interim analysis, both were re-estimated and the initial sample size was adjusted based on the formula given by Chang [11]:

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

where E_0 and E_i are the pretrial and the interim observed effect sizes, and a is a tuning parameter 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, this does not preserve the type I error, so adjustment to the α level is required. In this study, the conservative Bonferroni correction method was used to adjust α level [13]. If a less conservative correction method had been chosen, the gain in power after SSR more likely would have been greater, but the relative comparisons would have been similar.

The pretrial 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 [14]. A single SSR was conducted at 6 months or 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, because in many chronic conditions where it takes some time to observe change 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 from baseline to the end of study between the treatment group and the placebo group; missing values were imputed using last observation carry forward because of its simplicity and the assumption of nondifferential dropout, and the longitudinal nature of the data [16]. The test was conducted at the end of the planned treatment period without adjustment for covariates. The secondary analysis method was the linear mixed effects model, which tested the difference in the slopes of the ADAS-Cog between treatment arms with adjustment for age and education [2]. For all analyses, the missing data pattern present in the meta-database was used to realistically simulate dropouts. Observations were missing in simulated data sets 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 .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 (Figs. 1 and 2). Approximately 25% of trials required at least a doubling of the sample size regardless of initial sample sizes. When the initial sample size was 50 per group, 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 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. When the initial sample size was more than 200, the gain in power was smaller regardless of the type of trials. When the effect size was smaller, the power before and after SSR also became smaller; but the gain in power actually increased over larger initial sample sizes (Fig. 1). Under the same SSR method, the longer trial duration did not generate larger gains in power (Fig. 3). 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 (Fig. 3).

The "effect size method" generally resulted in greater gain in power than the "variance only method" (Fig. 2). 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 for Wilcoxon tests are similar to those for linear mixed effects model tests. Simulation results with treatment effects generated from normal distributions were

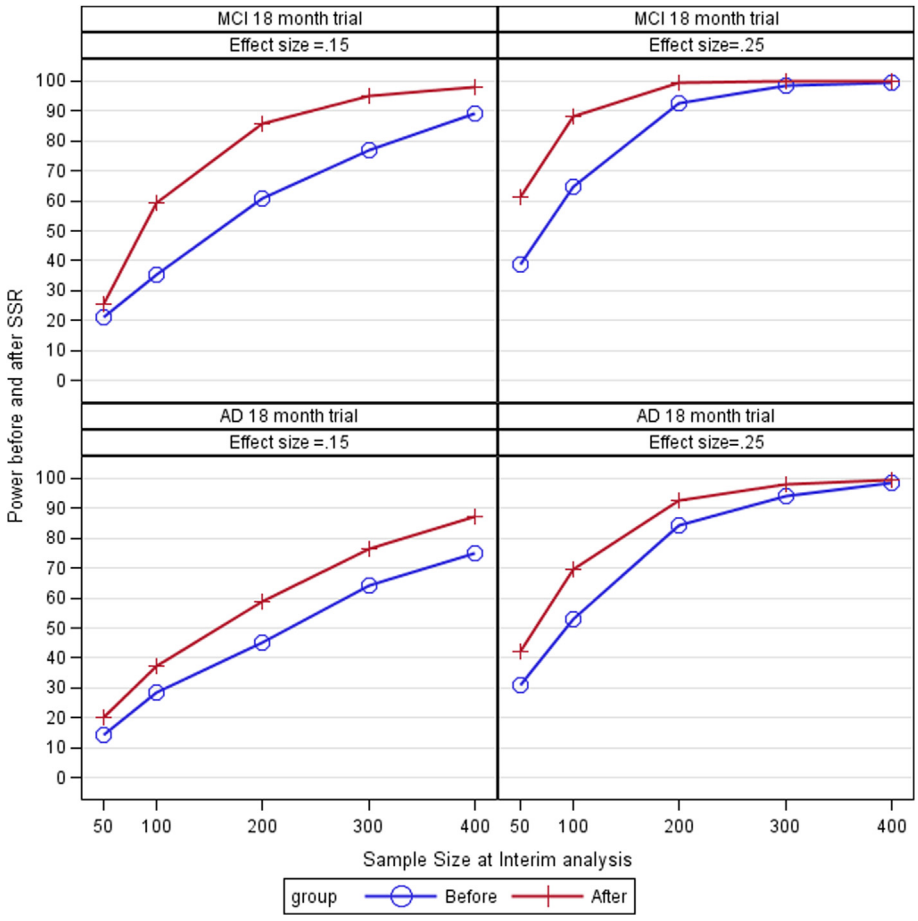

Fig. 1. Power comparison before and after sample size re-estimation (SSR) based on variances at 6 months. SSR results in gain in power for small to medium effect sizes from both Alzheimer's disease (AD) and mild cognitive impairment (MCI) trials.

compared with those from χ^2 distributions and the differences were very small (data not shown).

4. Discussion

Based on our simulations, in terms of power improvement, SSR adaptive designs can be effective for trials with small or medium initial sample sizes in AD and MCI. The effectiveness depends on factors such as the number of subjects accumulated for the interim analysis, the true treatment effect size, and the type of uncertainty in the pretrial estimates (effect sizes or variances). Too few subjects accumulated for the interim analysis might lead to imprecise estimates of treatment effects or variances, thus resulting in the poor prediction of sample size adjustments; too many subjects, subjects are already enrolled and trials without SSR already have adequate power. The smaller the true treatment effect, the more subjects are needed at the interim analysis to obtain precise estimates.

It should be noted in our simulations that even after SSR, the power for trials with small or medium initial sample sizes is still not optimal. However, the gain in power is substantial relative to that in the study as planned. Considering that these small or medium sample sizes are often used in feasibility studies or phase 2 studies, missing a potentially effective drug at that stage could be a major mistake. The application of SSR could come to the rescue. Although the gain in power is relatively small for trials with large initial sample sizes; SSR can still be useful for early termination for futility. For example, in the Vitamins B (HC) trial, at 12 months, the mean difference in the change of ADAS-Cog from baseline between treatment arms is 0.3 with a sample standard error of 6.1, leading to an effect size of 0.05; compared with the hypothesized effect size 0.15, it is small enough to open seriously the discussions of the termination of the ongoing trial for futility. Otherwise, the sample size has to be increased to $(.15/0.05)2*409 = 3681$ so that the initial power can be achieved. Whether the trial would produce an undetectable effect size (i.e. the drug works with less than planned effectiveness) or the trial would have been deemed futile if the interim analyses had been done, it is as much a clinical decision as a statistical one, but had an SSR process been in place, the decisions could have been considered.

Although the uncertainty in the pretrial estimates determines the SSR method, the "variance only" method is preferred over the "effect size" method [17,18], which

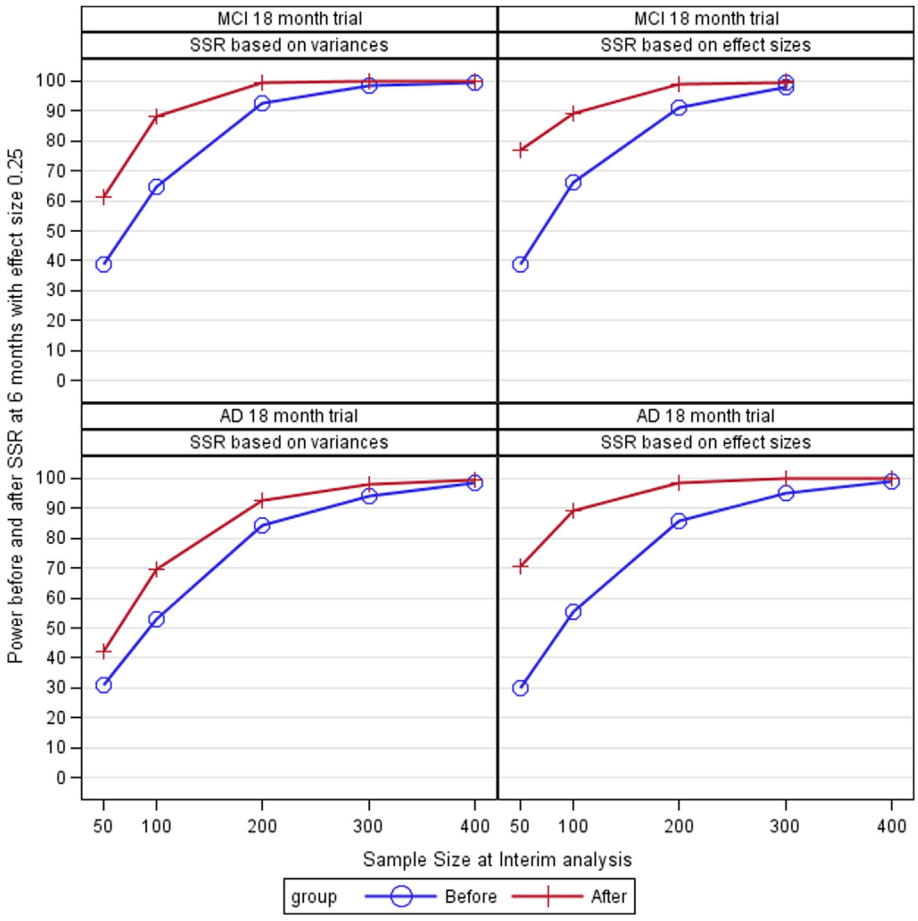

Fig. 2. Comparison between sample size re-estimation (SSR) at 6 months based on variances and based on effect sizes. Although both SSR methods led to gain in power, SSR based on effect sizes led to larger gain, especially for small to medium sample sizes.

emphasizes the importance of pretrial estimate of the treatment effect. 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.

Finally, it has been demonstrated that group sequential designs (GSDs) are more efficient than adaptive designs when the treatment effect needs re-estimation [19,20]. However, these studies of GSDs have assumed the variance of the outcome is known and fixed; whereas in SSR as applied in our study, this assumption is not required. Furthermore, the better efficiency in GSDs is at the price of "a large number of interim analyses, a large up-front sample size commitment and aggressive early stopping boundaries" [21]. Thus, although GSDs are theoretically interesting, they may not be pragmatic in AD trials. The purpose of those GSDs is also different from that of the SSR adaptive designs. The former imposes a fixed maximum sample size and aims mainly to stop the trial earlier for futility or efficacy with well-known limitations for safety and subgroup evaluations and potentially biased efficacy assessments. The latter allows a flexible maximum sample size and aims primarily to expand the study [22]. Therefore, the investigation of SSR based on both variances and effect sizes are useful.

For longitudinal studies, longer trials naturally have more power for effective treatments. However, our simulation indicated little difference in power between 18 months and 24 months trials after SSR. One explanation is the relatively small treatment effect was not enough to overcome the heterogeneity and inconsistency in ADAS-Cog within 6 months. This would also explain the lack of differences between SSR at 6 months and 12 months. An alternative would be to measure more frequently and use more measurements at the interim analysis to estimate the variances or the effect sizes.

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 RCT designs because RCT designs already offer adequate power. This is in contrast to the results of many completed clinical trials with equally large or even larger sample sizes [14,23]. The reason may be that, in our simulations, moderate treatment effects were assumed to exist at each measurement and persist from the beginning to the end of the study; although in the completed trials, the observed treatment effects were smaller and inconsistent over time. This contrast might indicate that if a moderate, clinically meaningful treatment effect indeed persists and can be reflected in the change of ADAS-Cog, a large sample would not be needed to detect it. However, in practice, the

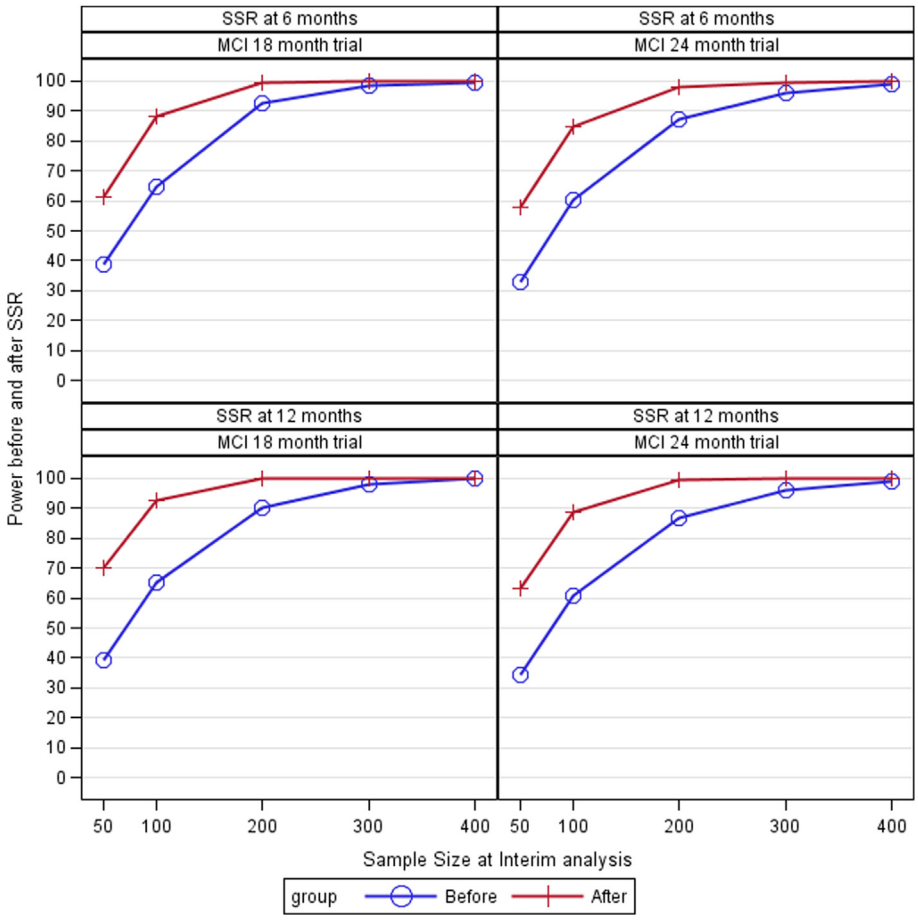

Fig. 3. Power comparison by trial duration and by the time of sample size re-estimation (SSR) based on variances. SSR at later time points (12 months) showed only slightly larger gain in power compared with SSR at earlier time points (6 months). There was also little gain in power with longer trials (24 months) compared with shorter trials (18 months).

large degree of uncertainty in effect sizes or variances prevents efficient trials with moderately large sample sizes.

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. Although we recommended SSR for trials with initial sample sizes less than 200 per arm, the optimal pretrial sample size was not determined.

Second, the possible impact of the recruitment rate on the time of SSR was not investigated. Very fast recruitment rates mean that at the interim analysis, most or even all 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 [24]. However, we recognize that stopping a trial involves much more than a statistical calculation.

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. Additionally, as the variances of outcome increase over time, the estimate of the variance at interim analyses may underestimate that of later time points.

Fourth, the flexibility to recruit additional subjects and the gain in power after SSR introduces complexity of logistics, masking treatment assignments, and statistical analysis.

Finally, we have not conducted a comprehensive evaluation of the predictability of SSR based on earlier outcomes at

Table 1
Increase in sample sizes after sample size re-estimation (SSR) by initial sample sizes

	Initial sample sizes				
	50	100	200	300	400
SSR method	Increase in sample sizes after SSR, mean (SD)				
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)

6 or 12 months, for later outcomes at 18 or 24 months. A brief investigation of the trials used for this study showed that the predictability is acceptable. For example, in trial HC, the effect size estimated at 12 months was 0.05, which was much less than the hypothesized effect size 0.15; and indeed at the end of the study, the effect size was −0.05, which was also less than the hypothesized 0.15. Similar results were concluded from the other trials as well. With negative trials this is certainly reasonable, but for trials with different therapeutic mechanisms of action and therapeutic lags, the predictability of successful outcomes from early measurements is harder to assess.

5. Conclusion

The SSR adaptive designs can improve efficiencies for AD and MCI trials. It can lead to significant gains in power for trials with small or medium initial sample sizes, or avoid the exposure of a large number of patients to ineffective treatments by stopping the trial earlier for futility. Considering the need to identify effective treatments, the continuous increase in sample size for AD trials, and the difficulty in estimating pretrial treatment effects, an SSR adaptive design can be a superior alternative to the typical RCT design.

Acknowledgments

Presented, in part, at Alzheimer's Association International Conference meeting, Boston, Massachusetts, July 2013.
Funding acknowledgments: Funding for this reported was provided by NIH R01 AG037561 (LSS, REK, GRC). Data used in the preparation of this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI, NIA U01 AG024904) database (www.loni.ucla.edu/ADNI), and from the ADCS (NIH AG10483).
Competing interests: Dr. Guoqiao Wang reports receiving grant support from NIA (R01 AG 037561).
Dr. Richard E. Kennedy reports receiving grant support from NIA (R01 AG 037561, R01 AG015062), NINDS (U01 NS41588), NHLBI (T32 HL072757), NIDDK (P60 DK079626), and the Department for Education (H133A070039).
Dr. Gary R. Cutter reports receiving grant or research support from Participation of Data and Safety Monitoring Committees: All the below organizations are focused on medical research: Apotek, Biogen-Idec, Cleveland Clinic, Glaxo Smith Klein Pharmaceuticals, Gilead Pharmaceuticals, Modigenetech/Prolor, Merck/Ono Pharmaceuticals, Merck, Neuren, Revalesio, Sanofi-Aventis, Teva, Vivus, NHLBI (Bone Marrow Transplant Protocol Review Committee), NINDS, NMSS, NICHD (OPRU oversight committee). Consulting, Speaking fees & Adviosry Boards: Alexion, Allozyne, Bayer, Celgene, Coronado Biosciences, Consortium of MS Centers (grant), Diogenix, Klein-Buendel Incorporated, Medimmune, Novartis, Nuron Biotech, Receptos, Spiniflex Pharmaceuticals, Teva pharmaceuticals. Dr. Cutter is employed by the University of Alabama at Birmingham and President of Pythagoras, Inc. a private consulting company located in Birmingham AL.
Dr. Lon S. Schneider reports being an editor on the Cochrane Collaboration Dementia and Cognitive Improvement Group, which oversees systematic reviews of drugs for cognitive impairment and dementia; receiving a grant from the Alzheimer's Association for a registry for dementia and cognitive impairment trials; within the past 3 years receiving grant or research support from NIA (R01 AG 037561), Baxter, Eli Lilly, Forum, Genentech, Lundbeck, Merck, Novartis, Pfizer and Tau Rx; and having served as a consultant for or receiving consulting fees from AC Immune, Allon, AstraZeneca, Avraham Pharmaceutical, Ltd, Baxter, Biogen Idec, Cerespir, Cytox, Elan, Eli Lilly, Forum, GlaxoSmithKline, Johnson & Johnson, Lundbeck, Merck, Pfizer, Roche, Servier, Takeda, Toyama, and Zinfandel.
Authors' contributions:
Guoqiao Wang, PhD: Drafting/revising the manuscript for content, including study concept or design, analysis and interpretation of data.
Richard E. Kennedy, MD, PhD: Drafting/revising the manuscript for content, including study concept and design, analysis and interpretation of data.
Gary R. Cutter, PhD: Drafting/Revising the manuscript for content, study concept and design, analysis and interpretation of data.
Lon S. Schneider, MD: Drafting/Revising the manuscript for content, including medical writing for content, study concept or design, analysis and interpretation of data.
Funding sources: Supported by NIH (R01 AG037561).

RESEARCH IN CONTEXT

1. Systematic review: We reviewed existing literature on clinical trial designs in Alzheimer's disease (AD). Several researchers have recommended that novel designs such as adaptive designs should be used to facilitate the development of effective treatments instead of the double-blind, placebo controlled, parallel group design. To provide guidance, we evaluated the effect of sample size re-estimation (SSR) in an adaptive trial designs for AD using simulation based on a meta-database of six AD clinical trials and observational studies.
2. Interpretation: Our results showed that SSR can be effective for AD clinical trials with small or medium initial sample sizes.
3. Future directions: Investigators should consider SSR adaptive designs under appropriate circumstances. Further research should evaluate the applicability of novel designs for specific populations.

References

[1] Thies W, Bleiler L. 2013 Alzheimer's disease facts and figures. Alzheimers Dement 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] Cummings J, Gould H, Zhong K. Advances in designs for Alzheimer's disease clinical trials. Am J Neurodegener Dis 2012;1:205.

[4] Chow SC, Chang M. Adaptive design methods in clinical trials—a review. Orphanet J Rare Dis 2008;3:11.

[5] Kennedy RE, Cutter GR, Schneider LS. Effect of *APOE* genotype status on targeted clinical trials outcomes and efficiency in dementia and mild cognitive impairment resulting from Alzheimer's disease. Alzheimers Dement 2014;10:349–59.

[6] Wechsler D. Wechsler memory scale-revised. San Antonio: Psychological Corporation; 1987.

[7] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med 2006;25:4279–92.

[8] Schneider LS, Kennedy RE, Cutter GR. Requiring an amyloid-β_{1-42} biomarker for prodromal Alzheimer's disease or mild cognitive impairment does not lead to more efficient clinical trials. Alzheimers Dement 2010;6:367–77.

[9] 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. N Engl J Med 2005;352:2379–88.

[10] 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.

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

[12] Lawrence Gould A, Shih WJ. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. Commun Stat Theory Methods 1992;21:2833–53.

[13] Dunn OJ. Multiple comparisons among means. J Am Stat Assoc 1961; 56:52–64.

[14] Schneider LS, Sano M. Current Alzheimer's disease clinical trials: methods and placebo outcomes. Alzheimers Dement 2009;5:388–97.

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

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

[17] Proschan MA. Sample size re-estimation in clinical trials. Biom J 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] Tsiatis AA, Mehta C. On the inefficiency of the adaptive design for monitoring clinical trials. Biometrika 2003;90:367–78.

[20] Jennison C, Turnbull BW. Efficient group sequential designs when there are several effect sizes under consideration. Stat Med 2006;25:917–32.

[21] Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: a practical guide with examples. Stat Med 2011; 30:3267–84.

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

[23] Knopman DS. Clinical trial design issues in mild to moderate Alzheimer disease. Cogn Behav Neurol 2008;21:197.

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

Supplementary Table 1
Placebo-controlled and observational studies included in this study

Study (code)	Design	N	Duration (months)
Selegiline, vitamin E (SL)	RCT, moderate to severe AD	341	24
Memory impairment study (MIS), donepezil, vitamin E	RCT, MCI	769	36
Simvastatin (LL)	RCT, mild to moderate AD	406	18
Vitamins B (HC)	RCT, mild to moderate AD	409	18
Docosahexaenoic acid (DHA)	RCT, mild to moderate AD	402	18
Alzheimer's Disease Neuroimaging Initiative (ADNI)	Observational, AD, MCI, normal	800	36 (AD) 48 (MCI) 48 (NL)

Abbreviations: RCT, randomized clinical trial; AD, Alzheimer's disease; MCI, mild cognitive impairment; NL, normal.

Supplementary Table 2
The mean difference (SD) in ADAS-Cog score between the placebo group and the treatment group in simulated trials

Study (code)	6 months	12 months	18 months	24 months
Selegiline, vitamin E (SL)	1.25 (1.29)	1.66 (1.64)	1.86 (1.84)	1.89 (2.21)
Memory impairment study (MIS) donepezil, vitamin E	1.02 (0.62)	1.10 (0.69)	1.25 (0.84)	1.39 (0.95)
Simvastatin (LL)	1.45 (1.22)	1.72 (1.37)	2.13 (1.53)	X
Vitamins B (HC)	1.33 (0.99)	1.54 (1.12)	2.01 (1.40)	X
Docosahexaenoic acid (DHA)	1.25 (1.11)	1.60 (1.40)	2.06 (1.65)	X
ADNI MCI (ADNI)	1.10 (0.60)	1.19 (0.69)	1.33 (0.88)	1.47 (1.00)
ADNI AD (ADNI)	1.25 (0.83)	1.53 (1.10)	X	2.24 (1.52)

NOTE. X means that there are no scheduled measurements at that time point.
NOTE. Sample size per group: 200, effect size: 0.25, number of simulated trials per scenario: 1000.