Sample Sizes for Randomized Controlled Trials Utilizing Bayesian Response Adaptive Randomization for Continuous Outcomes

Vahan Aslanyan1, BA; Michelle Nuño1,2, PhD; Trevor Pickering1, PhD; Lindsay A. Renfro1,2, PhD; Judy Pa3, PhD; & Wendy J. Mack1, PhD

1Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA

2Children's Oncology Group, Monrovia, CA

3Alzheimer’s Disease Cooperative Study (ADCS), Department of Neurosciences, University of California, San Diego, CA

Corresponding Author:

Wendy J. Mack, PhD

Department of Population and Public Health Sciences

University of Southern California

1845 N. Soto St.

Los Angeles, CA 90033

323-442-7200 (phone)

[wmack@usc.edu](mailto:wmack@usc.edu) (email)

**Abstract (250 words max)**

This paper describes a Bayesian approach for sample size estimation for multi-arm randomized controlled trials utilizing response adaptive randomization for continuous outcomes. Assuming normally distributed treatment effects and unknown but common treatment variance, this approach incorporates data from participants at prespecified points to re-estimate the total number of participants needed and change allocation proportions to favor treatments with larger treatment effect sizes and to update prior assumptions of the treatment parameters. The estimated sample size should be sufficient to show that either one of the treatment effect sizes is greater than 0, or that all treatment effect sizes are smaller than a clinically meaningful treatment threshold at prespecified probabilities. Using hypothetical scenarios, this approach shows that two interim analyses, conducted earlier in the course of the trial may result in less overall number of participants compared with sample size calculation approaches based on hypothesis testing. Incorporating response adaptive randomization, especially for cases with distinctly different treatment effects, results in an unbalanced allocation of participants to the arms, and the sample size increases to compensate for the imbalance. This is also illustrated using characteristics of two concluded randomized controlled trials. This study concludes that early and frequent interim analyses may reduce the number of participants needed for a conclusive trial for trials with constant participant allocations. The ethical benefits of response adaptive randomization should be considered against the economic benefits of constant allocation when designing a trial incorporating this method.

**Introduction**

Randomized controlled trials (RCT) are the gold standard for testing efficacy and safety outcomes of treatments in populations of interest. The conduct of RCTs relies on the equipoise principle, which states that *a priori* all proposed treatments are believed equally likely to be effective1. Equipoise serves as the basis for null hypothesis testing, where study researchers hypothesize that all candidate treatments have the same treatment effect (or that treatment effect differences are zero). Type I and Type II errors are key considerations when designing an RCT. Type I error is the probability of falsely rejecting of the null hypothesis, while Type II error is the probability of rejecting the true alternative hypothesis. Power is the probability of avoiding a Type II error and is closely related to the sample size of the study2. During the design stage, a sample size is determined such that Type I and II errors are within a prespecified bounds while also accounting for the variance of measurements and for the magnitude of a clinically relevant difference between treatments. For many scenarios (i.e. independent continuous/binary, clustered continuous/binary, repeated measurement outcomes, time-to-event outcomes), closed form solutions for sample size determination under frequentist approach have been derived and are widely implemented in RCTs3.

Most sample size estimation techniques assume that the allocation of participants to treatment arms remains constant throughout the trial. Response adaptive randomization (RAR) dynamically adjusts the allocation of participants based on accumulated interim data such that participants are more likely to be allocated to better performing treatment arms. Bayesian RAR incorporates initial assumptions about the parameters of interest by specifying prior distributions for these parameters and using them to calculate posterior distributions once the data becomes available. The posterior distribution parameters are then used to change the allocation probabilities for future participants (Aslanyan et al, under review). Frequentist sample size calculations are inappropriate in this scenario, since they fail to incorporate information from posterior distributions of parameters. Additionally, multiple interim analyses increase the probability of false discovery, requiring a strict control over type I error under frequentist framework. Bayesian sample size estimation methods can be implemented as alternatives to frequentist methods to avoid such limitations. They may also be seamlessly incorporated into RAR designs, such that at each interim analysis both the allocation and the required number of participants is evaluated, and changes are implemented at the same time.

Bayesian sample size estimation has been described for a single sample of binary endpoints4, two samples of binary endpoints (with re-estimation)5,6, time-to-event endpoints7, and for multiple samples of continuous endpoints4,8. Studies have demonstrated how to extend the approach for two samples of continuous endpoints to include re-estimation of sample size at interim analyses, assuming a constant allocation ratio throughout the trial9. This study also highlighted the need for further examination of the method with multiple sequential interim analyses and when allocation is not 1:1. Additionally, efforts for sample size estimation (and re-estimation) have been described for RARs based on significance testing10, but not for Bayesian RARs, indicating a need for studies describing and benchmarking sample size estimation methods for Bayesian RARs.

In this study, we describe a sample size determination method for multiple treatment Bayesian RAR RCTs with continuous outcomes. We then assess the performance of this method with the performance of hypothesis testing based approach, Bayesian sample size estimation with no interim analyses, with interim analyses and constant allocation ratios. The estimated sample size should be sufficient to ensure that either one of the treatments is superior to the control or that all treatments are not better than control by some clinically relevant treatment effect in all cases.

**Description of Bayesian Response Adaptive Randomization**

In general, the allocation probability of a participant to treatment i is determined based on the posterior probabilities estimating which treatment has the highest treatment effect:

δi is the treatment effect for treatment i, the parameter c in ri is a tuning parameter and is set to be n/2N, where n is the number of participants who have currently completed the trial, and N is the target sample size. The parameter will start at 0 at the beginning of the recruitment (n=0), ensuring an initial run-in period with equal randomization, and will gradually increase to ½ when n=N.

**Description of Bayesian approach to sample size determination**

This paper follows conventions of previously developed methods8,9. To summarize, the patients are randomized between a control treatment E0 and k experimental treatments E1,…,Ek. Outcome yij of the i-th patient on treatment Ej follows a normal distribution with mean µj and precision υ (precision is the inverse of variance and assumed to be the same for all treatments) for i=1,…,nj (nj is Nri where N is total sample size), and j=1,…k. Independent normal priors are suggested for µj, and gamma prior is suggested for υ.

These priors lead to independent normal posteriors for µj and υ.

where (, , and . Here is the mean response of participants on Ej, and . Treatment effects, denoted by δj will then follow a prior normal distribution with mean µj- µ0 and precision υ, and posterior distribution with mean µ1j- µ10 and precision D1jυ, where D1j=q1jq10/( q1j+q10). In this study, treatments are assumed to have the same variance, equal to the maximum posterior variance of the treatment effects, so all D1j-s are calculated and then the common D1 is set to the minimum of D1j-s. The number of recruited participants should be sufficient to ensure that none of the treatments have clinically important effect (δ\*), or that at least one treatment shows promise at prespecified probabilities. In other words,

Where η and ζ are probability thresholds. Previous studies describing these methods4,8,9 have shown that suitable sample sizes need to ensure that the above condition is satisfied with a probability ξ when

where n is the total sample size, is the 100ξ% point of of beta distribution with parameters 0.5n and α0, and and are 100ζ% and 100η% points of t distribution with 2α1 degrees of freedom.

This method can be extended to facilitate interim sample size re-estimation9. At each interim stage, distributions of means and precision are updated with the collected data. These updated distributions are then used to re-estimate total number of participants required to satisfy the condition above. Additionally, in RAR RCTs, the allocation vector r will also be updated at the interim stage. This new vector will be used to calculate the necessary sample size while intentionally skewing assignment towards better performing treatment arms.

To assess the performance of Bayesian sample size re-estimation for RAR, hypothetical scenarios will be considered. To reflect Type I and II error parameters in frequentist design, η will be set to 0.95 (reflecting Type I error rate=0.05), ζ=0.9 (reflecting power=0.9 (Type II error rate=0.1)), and ξ=0.95 (95% chance of a conclusive trial). One placebo and three active treatment arms will be considered with treatment effect sizes of 0, 0.1, 0.2, and 0.6 (precision=1). Three different clinically relevant treatment effect thresholds (δ\*) will be considered (0.1, 0.2, 0.3). Additionally sample size will be re-estimated at either one interim stage at 12.5% (25%, 50%) enrollment or two interim stages at 25% and 50% (25% and 75%) enrollment. Full sets of parameters are presented in Table 1. Priors for precision were sampled from a wide range of values of α0 (5000 randomly selected values from Uniform (0,20) distribution).

The relative performance of the method will also be presented for real life examples of Lecanemab Phase III study11, with one active arm; and the ARGO study, a Phase II investigation of Tideglusib with three active arms12. The first study was successful, while the second study did not demonstrate efficacy in treating patients.

The code for this project was written in R (version 4.2.3). Code will be available on GitHub. Sample size estimation based on hypothesis testing was conducted in PASS 2023.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scenario** | **Treatment Effect Sizes** | **Clinically Important Effect Size** | **Number of Interim Analyses** | **Spacing of Interim Analyses (% Enrollment)** |
| 1 | 0, 0.1, 0.2, 0.6 | 0.1 | 1 | 50 |
| 2 | 25 |
| 3 | 12.5 |
| 4 | 2 | 25, 50 |
| 5 | 25, 75 |
| 6 | 0.2 | 1 | 50 |
| 7 | 25 |
| 8 | 12.5 |
| 9 | 2 | 25, 50 |
| 10 | 25, 75 |
| 11 | 0.3 | 1 | 50 |
| 12 | 25 |
| 13 | 12.5 |
| 14 | 2 | 25, 50 |
| 15 | 25, 75 |
| 16 | 0, 0.45 | 0.373 | 1 | 50 |
| 17 | 25 |
| 18 | 12.5 |
| 19 | 2 | 25, 50 |
| 20 | 25, 75 |
| 21 | 0.25, 0.03, 0.21, 0.35 | 0.64 | 1 | 50 |
| 22 | 25 |
| 23 | 12.5 |
| 24 | 2 | 25, 50 |
| 25 | 25, 75 |

**Table 1: Scenarios under consideration.** All studies have used uninformative priors for μ-s (0-s) and q-s (1-s). Scenarios 16-20 use data from Lecanemab Phase III study, 21-25 use data from the ARGO study.

**Hypothetical Scenarios**

For the three clinically important effect sizes, estimating sample size using Bayesian approach results in sample sizes that are considerably bigger compared to the numbers estimated using hypothesis testing approaches. The sample size estimation with hypothesis testing approach assumes that the precision is known, while the Bayesian approach does not make such an assumption. In case the difference between the observed and expected precisions is not big, hypothesis testing approach does result in an adequately powered study; however, with Bayesian approach a study recruits more participants to account for the variance in the estimate of precision and reduces the possibility of an underpowered analysis if the observed and expected variances are different from each other.

Within Bayesian approach, incorporating interim analyses in the design is shown to reduce the variance of precision (βposterior increases as more data becomes available and the variance of a gamma distribution with parameters α and β is α/β2), reducing the number of participants needed for a conclusive trial. The simulations also show that interim analyses as early as at 12.5% target enrollment may provide enough data to reduce the number of participants, showing an economic advantage (Scenario 3, 8, 13). Additionally, increasing the number of interim analyses increases the number of updates of prior distributions, leading to more data-guided distributions and thus to less participants needed for a conclusive trial compared to scenarios with one interim analysis only, regardless of the timing of the single interim analysis (Table 1).

Incorporating RAR in the sample size reestimation generally results in group imbalances when the treatment effects are different from each other (as is the case in scenarios 1-15). RAR is dependent on collected data per treatment arm and the variance within the arm, thus earlier analyses may result in estimates with bigger variances and bigger imbalances (Scenarios 1-3, 6-8, 11-13). The second interim analysis decreases the number of participants needed in these cases too, indicating that RCTs with RAR and Bayesian sample size reestimation designs may require less participants for a conclusive trial compared to RCTs designed on hypothesis testing approach and no interim analysis. However, the imbalance that RAR introduces makes it a less economically attractive option compared to constant allocation, thus the ethical benefits that RAR may provide should be considered against potential savings of resources, especially if the probability of a futile trial is high.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scenario** | **Sample Size—Hypothesis Testing** | **Sample Size w/o interim Analysis** | **Sample Size w Interim Analysis** | **Sample Size with RAR** |
| 1 | 2102/2102/2102/2102 | 4065/4065/4065/4065 | 2215/2215/2215/2215 | 2161/2279/2289/2289 |
| 2 | 1861/1861/1861/1861 | 1659/2162/2296/2302 |
| 3 | 1808/1808/1808/1808 | 1583/2222/2585/2659 |
| 4 | 1005/1005/1005/1005 | 878/1173/1245/1248 |
| 5 | 1396/1396/1396/1396 | 1242/1623/ 1718/1723 |
| 6 | 526/526/526/526 | 1257/1257/1257/1257 | 677/677/677/677 | 668/690/703/707 |
| 7 | 532/532/532/532 | 512/583/667/710 |
| 8 | 493/493/493/493 | 515/567/706/871 |
| 9 | 281/281/281/281 | 267/308/354/377 |
| 10 | 399/399/399/399 | 387/438/503/539 |
| 11 | 235/235/235/235 | 675/675/675/675 | 361/361/361/361 | 359/366/374/380 |
| 12 | 270/270/270/270 | 274/290/329/379 |
| 13 | 240/240/240/240 | 286/272/336/465 |
| 14 | 140/140/140/140 | 145/149/175/205 |
| 15 | 202/202/202/202 | 209/217/251/290 |
| 16 | 626/626 | 1658/1658 | 1017/1017 | 871/1612 |
| 17 | 840/840 | 626/4457 |
| 18 | 786/786 | 588/3599 |
| 19 | 493/493 | 610/735 |
| 20 | 630/630 | 586/686 |
| 21 | 56/28/56/56 | 331/165/331/331 | 171/85/171/171 | 176/85/173/178 |
| 22 | 198/99/198/198 | 284/100/231/275 |
| 23 | 219/110/219/219 | 319/119/250/289 |
| 24 | 58/29/58/58 | 64/8/40/50 |
| 25 | 87/43/87/87 | 186/44/120/167 |

**Table 2: Estimated total sample sizes for scenarios under consideration.** For scenarios 1-15, sample sizes were estimated using multi-arm tests for the difference between treatment and control means assuming equal variance. For scenarios 16-25, two sample t-tests were used per the respective study’s protocol (Lecanemab (Scenarios 16-20), and Tideglusib (Scenarios 21-25)

**Case study—Lecanemab**

An 18-month, multicenter, double blind, Phase III trial tested the efficacy of Lecanemab against a placebo in persons with early Alzheimer’s disease with a primary end-point of Clinical Dementia Rating-Sum of Boxes (CDR-SB)11. The sample size for this study was estimated using data from the respective Phase II study, with estimated treatment difference of 0.373, and estimated standard deviation of 2.031 (variance=4.12, ν0=0.24) for the change from baseline in the placebo arm. The study aimed to recruit 1566 subjects (equal randomization) to have 90% power to detect differences between Lecanemab and placebo using a 2-sample t-test at Type I error rate=0.05 and assuming 20% dropout. In the modified intent-to-treat population, adjusted mean change between groups was -0.45 (95% CI: (-0.67, to -0.23) for the primary endpoint11 (N=859 in Lecanemab group, 875 in placebo). The estimated standard error of change is thus 0.1122, corresponding to standard deviation of 2.34.

This trial showed that Lecanemab was superior to placebo, significantly reducing cognitive decline measured by CDR-SB. The relative efficacy of Lecanemab was reflected in the interim sample size re-estimation with equal and response-adaptive randomizations (Table 2, Scenarios 16-20). The number of required participants reduced from initial assessment (3316) for all scenarios. RAR produced big group imbalances (favoring Lecanemab and thus allocating more participants to the treatment arm), which subsequently increased the number of required participants to accommodate such an imbalance when performing only one interim analysis. As expected, increasing the number of interim analyses incorporates more treatment data, resulting in less group imbalance and smaller sample sizes compared with cases with one interim analysis only. We see that two interim analyses at 25/50% target enrollment with equal randomization required less participants to achieve the sample desired characteristics as the sample size estimated by t-test (1305 total assuming no dropout), again indicating that the earlier the sample size is reestimated, the less participants the study would need to test efficacy. This trend is also observed for combined RAR and sample size reestimation design, with the caveat that enrolling more patients in the better performing arms requires more participants enrolled in the study too.

**Case study—Tideglusib**

A 26-week, multicenter, double blind, Phase II trial tested the efficacy of Tideglusib at three different dosage intakes (500mg/day, 1000mg/day once a day, 1000mg/day every other day) against a placebo in persons with mild to moderate Alzheimer’s disease with a primary endpoint of ADAS-cog15. The sample size for this study was estimated using data from a pilot study, where a clinically relevant difference between active and placebo arms of 4.7 (SD=7.3, variance is) was reported for the primary endpoint. Assuming a 30% drop-out rate, the study aimed to recruit 280 patients ( with 1/7of participants allocated to the 500mg/day arm, and 6/7 of participants allocated to the other arms equally) to provide 90% statistical power to detect a 4.5 point difference in the ADAS-cog15 between each treatment and placebo group using t-test at Type I error rate=0.0512. In the intent-to-treat population, mean change in 82 participants in the placebo arm had mean 1.7 ( SD=6.9); mean change in 43 participants receiving Tideglusib at 500mg/day had mean 0.2 (SD=6.7); mean change in 85 participants randomized to 1000mg/day every other day was 1.4 (SD=6.6); and the mean change among 84 participants receiving Tideglusib at 1000mg/day every day was 2.6 (SD=7.4)12. The prior variance is thus 53.29, ν0= 0.019.

This study concluded that Tideglusib produced no clinical benefit. This is reflected in 4 out of 5 scenarios for this intervention, where most participants were assigned to the placebo arm for the combined RAR and sample size reestimation design. Additionally, in all 5 scenarios the least number of participants were assigned to the arm with no treatment difference. Similar to Lecanemab scenarios, the number of required participants decreased as interim analyses were incorporated into the design, however, interim analyses at earlier recruitment targets did not result in overall less participants enrolled, indicating the relative inefficiency of the interventions under consideration (Table 2, Scenarios 21-25). In scenarios 24 and 25, smaller number of required participants was needed to achieve a conclusive trial with 95% probability, reinforcing the need for multiple interim analyses while conducting RCTs incorporating Bayesian sample size reestimation in their designs.

**Conclusions**

This study describes a sample size determination method for multiple treatment Bayesian RAR RCTs with continuous outcomes. The study shows that early and multiple interim analyses in the study design result in less overall number of participants needed for a conclusive trial compared to the sample size based on hypothesis testing with no interim analyses, and that RAR-introduced imbalances would require more participants compared with constant allocation.

This study also demonstrates the method based on two completed RCTs—Lecanemab Phase III study and Tideglusib Phase II study. In these real-life scenarios, the conclusions from the hypothetical scenarios did not change much. In fact, for the Lecanemab study, simulations showed that changing design to estimate sample size on Bayesian approach and two interim analyses at 25 and 50% target enrollment and constant allocation may have reduced the number of participants needed in each arm by 133, resulting in significant savings of resources dedicated to the trial. In Tideglusib study, where the treatment effect sizes were close to each other, the RAR approach at 25 and 50% target enrollment correctly stopped recruiting participants into the arm with no change. This scenario also resulted in the least number of participants needed for a conclusive trial. Additionally, the allocation ratio clearly favoring the placebo arm may have indicated an earlier need to stop the trial due to futility.

The FDA guidance about adaptive designs for clinical trials of drugs and biologics13 encourages the use of adaptive designs in early phase exploratory trials. The design proposed in this study is highly suitable for exploratory trials. The advantages of Bayesian response adaptive randomization and sample size re-estimation can be used jointly to move candidate treatments to confirmatory stages, where hypothesis testing based designs and rigorous and conservative statistical analyses confirm efficacy of candidate treatments to show evidence of clinically meaningful differences.

**References**

1. Freedman B. EQUIPOISE AND THE ETHICS OF CLINICAL RESEARCH. *N Engl J Med*. 1987;317(3).

2. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. *Emerg Med J*. 2003;20(5):453-458. doi:10.1136/emj.20.5.453

3. Ahn C, Heo M, Zhang S. Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research.

4. Whitehead J, Valdés‐Márquez E, Johnson P, Graham G. Bayesian sample size for exploratory clinical trials incorporating historical data. *Statistics in Medicine*. 2008;27(13):2307-2327. doi:10.1002/sim.3140

5. Zhong W, Koopmeiners JS, Carlin BP. A two-stage Bayesian design with sample size reestimation and subgroup analysis for phase II binary response trials. *Contemporary Clinical Trials*. 2013;36(2):587-596. doi:10.1016/j.cct.2013.03.011

6. Pezeshk H, Gittins J. A Fully Bayesian Approach to Calculating Sample Sizes for Clinical Trials with Binary Responses. *Drug Information J*. 2002;36(1):143-150. doi:10.1177/009286150203600118

7. Cotterill A, Whitehead J. Bayesian methods for setting sample sizes and choosing allocation ratios in phase II clinical trials with time‐to‐event endpoints. *Statistics in Medicine*. 2015;34(11):1889-1903. doi:10.1002/sim.6426

8. Whitehead J, Cleary F, Turner A. Bayesian sample sizes for exploratory clinical trials comparing multiple experimental treatments with a control. *Statistics in Medicine*. 2015;34(12):2048-2061. doi:10.1002/sim.6469

9. Brakenhoff T, Roes K, Nikolakopoulos S. Bayesian sample size re-estimation using power priors. *Stat Methods Med Res*. 2019;28(6):1664-1675. doi:10.1177/0962280218772315

10. Li X, Hu F. Sample size re‐estimation for response‐adaptive randomized clinical trials. *Pharmaceutical Statistics*. 2022;21(5):1058-1073. doi:10.1002/pst.2199

11. van Dyck CH, Swanson CJ, Aisen P, et al. Lecanemab in Early Alzheimer’s Disease. *N Engl J Med*. 2023;388(1):9-21. doi:10.1056/NEJMoa2212948

12. for the ARGO investigators, Lovestone S, Boada M, et al. A Phase II Trial of Tideglusib in Alzheimer’s Disease. *JAD*. 2015;45(1):75-88. doi:10.3233/JAD-141959

13. FDA/CDER/"Bent R. Adaptive Designs for Clinical Trials of Drugs and Biologics.