Sample Size of Randomized Clinical Trial: A Simulation Study

Feifang Hu, Edison Choe and Lu Xia

Department of Statistics, University of Virginia,
Charlottesville, VA 22904, USA
fh6e@virginia.edu

Abstract. Randomization procedures have been extensively used in clinical trials and other experiments. In clinical trials, randomization tends to produce study groups comparable with respect to known and unknown covariates. It also removes investigators' bias in the allocation of patients. To implement adaptive randomization procedures in practice, we need to calculate their corresponding sample sizes. In the literature, sample sizes are calculated by ignoring the randomness of allocation. However, given a fixed sample size of a randomized design, the number of patients assigned to each treatment is a random variable and so is the power of the randomized design. Hence, under the sample sizes calculated by ignoring the randomness of the allocation, there is great possibility that we may not achieve the predetermined power we want. The central point of this paper is that the randomness of allocation in randomized designs should not be ignored, as doing so will significantly underestimate the sample sizes. In this paper, we focus on the sample sizes of two-arm (drug versus control) randomized clinical trials. First, we state the random power function for any given sample size and the properties of this power function. Based on the power function, we can calculate theoretical sample sizes for randomized designs. We then develop a simulation method to calculate sample sizes for randomized designs. The two methods are compared by some simulation studies.

Keywords. Randomization; Randomized designs; Restricted Randomization; Response-Adaptive Randomizations; Biased coin designs; Sample Size; Power; Sample Size Estimation.

1 INTRODUCTION

Randomization is the crux of valid experimental designs. By randomly allocating patients to either the treatment or the control group, the experimenter can negate or "average out" differences in individual patient characteristics that may affect the results and effectively prevent potential bias of both the investigators and the patients from the allocation process. And so by constructing comparable groups and eliminating bias, any differences between the treatment and the control groups at the end of the experiment can be confidently attributed to the treatment alone, ensuring the validity of the statistical tests and results. Rosenberger and Lachin (2002) and Hu and Rosenberger (2006) give a comprehensive review and study of randomization in clinical trials.

Given an experimental design, the investigator must determine an adequate sample size that meets the minimum predetermined power. Power is the probability of detecting a statistically significant difference between the treatment and the control groups if there is a difference indeed and power is directly proportional to sample size. This means that given all other things are constant, increasing the sample size will increase the power. In the interest of saving time and resources, the investigator normally calculates the minimum sample size required to achieve the desired power. Until recently in the literature, sample sizes for randomized designs have been calculated by simply assuming a fixed or predetermined allocation in each group. Although doing so will produce reasonable approximations, the sample sizes will almost always be underestimated. This is because the random factor in allocation contributes additional variability that should not be overlooked (Hu and Rosenberger, 2003). Therefore, a valid and more accurate calculation of the sample size for a randomized design should take into account the random factor of the patient allocation, which is the central point of this paper. A more detailed explanation will follow, and the application and evidence of this will be illustrated in the context of various randomized designs.

This paper focuses on three of five categories of randomized designs classified by Hu and Rosenberger (2006): the complete randomization (CR) design, the restricted randomization designs (Wei's Urn Design (UD) and the Generalized Biased Coin Design (GBC) are discussed here), and the response-adaptive randomization procedures. Within the last category is the Doubly Adaptive Biased Coin Design (DBCD), initially proposed by Eisele and Woodroofe (1995) and further explored by Hu and Zhang (2004). Calculations of sample size and power will be demonstrated for each of these designs using formulae described in Hu and Rosenberger (2006).

2 DEFINING POWER AND SAMPLE SIZE OF RANDOMIZED DESIGNS

2.1 Defining parameters and power

Here, we first define the parameters and briefly review the power of comparing two treatment groups (experimental and control) to provide a framework for discussing sample size. We use the same notation as in Hu (2006). Let n_E and n_C be the number of patients allocated to experimental and control groups, respectively (where n_E + n_C = n is the total sample size). We then define $X_1, ..., X_{n_E}$ as the responses of patients in the experimental group and $Y_1, ..., Y_{n_C}$ as the responses of patients in the control group, where

$$X_1 \sim N(\mu_E, \sigma_E^2) \text{ and } Y_1 \sim N(\mu_C, \sigma_C^2),$$

Here, (μ_E, μ_C, σ_E^2,σ_C^2) are unknown parameters.

Power, as stated before, is the ability to correctly identify a significant difference between the experimental and the control groups. In statistical terms, power is the probability of rejecting the null hypothesis (H_0) given that the alternative hypothesis (H_1) is true. For the purposes of this paper, we use a one-sided hypothesis testing as the following:

$$H_0 : \mu_E = \mu_C, \text{ v.s. } H_1 : \mu_E > \mu_C.$$

Given these definitions, we reject the H_0 at significance level α if

$$\frac{\hat{\mu}_E - \hat{\mu}_C}{\sqrt{\hat{\sigma}_E^2/n_E + \hat{\sigma}_C^2/n_C}} > z_{(\alpha)}, \tag{1}$$

where $\hat{\mu}_E, \hat{\mu}_C, \hat{\sigma}_E^2$, and $\hat{\sigma}_C^2$ are all estimators of the true parameters, $z_{(\alpha)}$ is such that $P(Z > z_{(\alpha)})$ = α, and Z is a standard normal random variable, meaning $Z \sim N(0, 1)$. Also let Ψ and ψ be the cumulative distribution function and density function of the standard normal distribution respectively. Therefore, the approximated power (β_0) can be expressed as

$$\beta_0 = P\left(\frac{\hat{\mu}_E - \hat{\mu}_C}{\sqrt{\hat{\sigma}_E^2/n_E + \hat{\sigma}_C^2/n_C}} > z_{(\alpha)}|H_1\right),$$

which can be shown to be approximately equivalent to

$$\beta_0 = P\left(z < \frac{\hat{\mu}_E - \hat{\mu}_C}{\sqrt{\hat{\sigma}_E^2/n_E + \hat{\sigma}_C^2/n_C}} - z_{(\alpha)}\right) \sim \Psi\left(\frac{\hat{\mu}_E - \hat{\mu}_C}{\sqrt{\hat{\sigma}_E^2/n_E + \hat{\sigma}_C^2/n_C}} - z_{(\alpha)}\right), \tag{2}$$

For the purposes of calculating power in this paper, $\hat{\sigma}_E^2, \hat{\sigma}_C^2$, and $\mu_E - \mu_C$ are assumed to be given, so the estimated parameters are not necessary. With a basic understanding of power, we can now move on to a discussion of sample size.

2.2 Determining sample size

Since power is directly proportional to sample size, as easily verified by the power function (2), this function can be used to derive sample sizes for randomized designs. Ignoring the randomness of allocation in randomized designs and assuming that n_E and n_C are fixed, the sample size (n_0) formula can be derived from the power function (2) as

$$n_0 = (\frac{\sigma_E^2}{\nu} + \frac{\sigma_C^2}{1-\nu})\frac{(z_{(\alpha)} + z_{(1-\beta)})^2}{(\mu_E - \mu_C)^2}, \tag{3}$$

where $0 < \nu < 1$ is the allocation proportion of $n_E/n \sim \nu$ for a given randomized design and n_0 is rounded up to the next integer.

In reality, however, n_E and n_C are random variables given a fixed n in randomized designs. Consequently, power is also a random variable. Taking this randomness into account, the approximated power function becomes

$$\beta_1 = \Psi\bigg(\frac{\mu_E - \mu_C}{\sqrt{\sigma_E^2/(\nu n + \tau\sqrt{n}Z) + \sigma_C^2/((1-\nu)n - \tau\sqrt{n}Z)}} - z_{(\alpha)}\bigg), \tag{4}$$

where τ is defined as the asymptotic result of

$$\sqrt{n}(n_E/n - \nu) \rightarrow N(0, \tau^2) \text{ with } \tau^2 > 0,$$

for a given randomized design. Note that the power function (4) is derived from (3) by replacing n_E and n_C with $\nu n + \tau\sqrt{n}Z$ and $(1-\nu)n - \tau\sqrt{n}Z$, respectively. Because the power function (4) is a random variable, it is more appropriate to use the approximated average power (β_2) given by

$$E(\beta_1) = \beta_2 = \int_{-\infty}^{\infty} \Psi\bigg(\frac{\mu_E - \mu_C}{\sqrt{\sigma_E^2/(\nu n + \tau\sqrt{n}x) + \sigma_C^2/((1-\nu)n - \tau\sqrt{n}x)}} - z_{(\alpha)}\bigg)\psi(x)dx. \tag{5}$$

In practice, however, β_2 is estimated by running a simulation of the power function given in (4). For each replication, Z is assigned with a randomly generated number from the standard normal distribution. Calculation of β_1 is replicated many times, and then the average is taken to estimate β_2. In this way, sample size (n_1) is found by finding the minimum n which produces an estimated β_2 that is greater than or equal to the predetermined power (β).

As explained, the nature of estimating β_2 assumes that the clinical trial or experiment is replicated many times. Practically speaking, however, the clinical trial is conducted only once. Thus, obtaining a specific power on average is hardly satisfactory. Instead, we want to be sure that β will be achieved with a certain high probability of $1 - \rho$. It can be shown that the smallest integer n that satisfies both

$$\frac{\sigma_E^2}{\nu n - z_{(\rho/2)}\tau\sqrt{n}} + \frac{\sigma_C^2}{(1-\nu)n + z_{(\rho/2)}\tau\sqrt{n}} < [\frac{\mu_E - \mu_C}{z_{(\alpha)} + z_{(1-\beta)}}]^2 \tag{6}$$

and

$$\frac{\sigma_E^2}{\nu n + z_{(\rho/2)}\tau\sqrt{n}} + \frac{\sigma_C^2}{(1-\nu)n - z_{(\rho/2)}\tau\sqrt{n}} < [\frac{\mu_E - \mu_C}{z_{(\alpha)} + z_{(1-\beta)}}]^2, \tag{7}$$

is the minimum sample size (n_2) required to achieve β with a probability of $1 - \rho$.

To recap, n_0 is the sample size given in (3), n_1 is the sample size minimizing the power function given in (5), and n_2 is the sample size minimizing equations (6) and (7). Application and comparison of these sample size formulae are illustrated for various randomized designs in the next section.

3 CALCULATING SAMPLE SIZE FOR VARIOUS RANDOMIZED DESIGNS

We start our discussion with complete randomization (CR), where each patient is randomly allocated to either the experiment or control group with a probability of 0.5. Now we show how to calculate the three different sample sizes: n_0, n_1, and n_2. For all calculations in this paper, we will use a significance level α of .05, set the requisite power β at .80, let $\mu_E - \mu_C = 1$ and $\sigma_E = 1$, and treat σ_C as an independent variable that is given. For a complete randomization procedure, $\mu = 1/2$ and $\tau = 1/2$.

(I). Calculation of n_0 (Take an example with $\sigma_C = 2$):

- Plug all the given parameters into equation (3), and here we get 61.85 as a result.
- Round up the result we get in the previous step to the nearest integer. Here, we get $n_0 = 62$.

(II). Calculation of n_1 (Take an example with $\sigma_C = 2$):

- We run simulations on the approximated power function (4) using statistical software. In this paper, we always use 10,000 replications for a simulation.

- Choose a guess for the sample size n by starting at n_0, in this case $n_0 = 62$, and plug all the given values into (4). Then repeatedly generate a random value for Z from the standard normal distribution for 10,000 times and calculate the approximated power (i.e. β_1) each time.
- Computing the mean of these 10,000 different values of β_1 will give an estimate for β_2. In our example, we get an estimate of .7962 for β_2.
- If the estimate is greater than or equal to 80%, we already get the n_1 we want. Otherwise, we need to increase n by one each time and repeat the above procedure until we get an estimate greater than or equal to 80%. In our example, since .7962 is less than 0.8, we run another simulation by increasing n by one to 63, and this time we get an estimate of .8016 for β_2, which exceeds 80%, so we stop. Therefore, $n_1 = 63$ in our example.

(III). Calculation of n_2 (Take an example with $1 - \rho = .90(\rho = .10)$):

- Choose a guess for the sample size n by starting at n_1, in this case $n_1 = 63$. We substitute in the previous given values and check whether the size n we choose will satisfy both equations (6) and (7). If it does, we already get the n_2 we want; otherwise, we have to increase n by one each time until it satisfies both equations (6) and (7). The last n we get in this procedure will be our n_2. In our example, we find that the minimum required $n = 72$, which is our n_2.

Now we consider restricted randomization, which includes Wei's Urn Design (UD)(Wei, 1977), and the Generalized Biased Coin Designs (GBC) (Smith, 1984). To compute n_0, n_1, and n_2 in restricted randomization, we use $\tau^2 = 1/12$ for UD and $\tau^2 = 1/44$ for GBC, and then follow the same processes as we have done in CR. For details about the two designs and the derivations about the parameters, please refer to Rosenberger and Hu (2004).

Finally, we consider the response-adaptive randomization designs. This paper focuses on a design called the Doubly-adaptive Biased Coin Design (DBCD), proposed by Eisele and Woodroofe (1995) and further developed by Hu and Zhang (2004). We also discuss a special case of DBCD called the Sequential Maximum Likelihood procedure (SMLE), introduced by Melfi and Page (2000). The details of ν and τ^2 can be found in Hu and Rosenberger (2006). We use the allocation function q $(\gamma > 0)$defined by Hu and Zhang (2004):

$$q^{(\gamma)}(x, y) = \frac{y(y/x)^\gamma}{y(y/x)^\gamma + (1-y)[(1-y)/(1-x)]^\gamma}, \ \gamma \geq 0. \tag{8}$$

4 SAMPLE SIZE RESULTS

4.1 Theoretical Results

All calculations were done using the statistical software R. For all procedures, sample sizes n_0, n_1, and n_2 have been calculated with all the given parameters and σ_C= 1, 2, and 4:

Table 1. Sample sizes of different randomize procedures ($\alpha = .05$, $\beta = .8$, $1 - \rho = .9$ and $\mu_E - \mu_C = 1$).

(σ_E, σ_C)	$(1,1)$	$(1,2)$	$(1,4)$	(σ_E, σ_C)	$(1,1)$	$(1,2)$	$(1,4)$
n_0	25	62	211	n_0(DBCD)	25	56	155
n_1 (CR)	26	63	212	n_1 (SMLE)	27	58	157
n_2 (CR)	28	72	233	n_2 (SMLE)	31	63	163
n_1 (UD)	26	63	211	n_1 (DBCD, $\gamma = 1$)	26	57	156
n_2 (UD)	26	68	223	n_2 (DBCD, $\gamma = 1$)	28	59	158
n_1 (GBC)	25	62	211	n_1 (DBCD, $\gamma = 4$)	26	57	156
n_2 (GBC)	25	65	217	n_2 (DBCD, $\gamma = 4$)	27	58	157

We could observe that although the difference between n_0 and n_1 may be negligible, the difference between n_0 and n_2 is not. When randomness of allocation is considered, a significantly greater sample size (n_2) is required to confidently achieve a given power than when randomness is ignored (n_0). We observe that both of the two response adaptive randomization procedures reduce the requisite sample size greatly from the required size given by CR and the Restricted Randomization.

4.2 Simulation Results

For the purpose of validating the theoretical calculations of n_1 and n_2, simulations of all the randomized designs described above have been performed using R and the procedures are described as following.

(I) Simulation of n_1:

- Choose a sample size n to start our simulation, usually n_0.
- Simulate the allocation of patients to the two treatment groups using the appropriate allocation procedure. For patients in the experimental group, generate values from $N(1,1)$ as treatment responses; for patients in the control group, generate values from $N(0,\sigma_C)$. After all the patients are being assigned, we know that there are n_E and n_C patients in the experimental and the control groups, respectively.
- Then, calculate the sample means and variances of the generated observations for both groups ($\hat{\mu}_E$, $\hat{\mu}_C$, $\hat{\sigma}_E^2$, $\hat{\sigma}_C^2$). Plug these values into equation (1) to determine whether or not H_0 is rejected.
- Replicate this whole process 10,000 times, and finally the power is calculated by dividing the number of times H_0 was rejected by 10,000.
- This simulation is run with different integer values of n until the minimum integer that produces a power of at least 80% is found. This value of n is then n_1.

(II) Simulation of n_2:

- The goal is to simulate the probability of having a power of at least 80% for a given randomized design and sample size.
- Simulation follows from what we have done for n_1, except that the number of successful rejections of H_0 is recorded for each possible value of n_E, which can range anywhere from 0 to n.
- For each individual case of n_E, divide the number of times H_0 was rejected in this case by the total number of times that this value of n_E occurred, and the result is the power.
- Finally, for all the cases of n_E that have a power of at least 80%, take the sum of all the times that these values of n_E occurred and divide it by 10,000. This value is the probability of having a power of at least 80%.
- We repeat this simulation with different values of n until we find a minimum integer that produces a confidence probability of at least 90%. This value of n is the n_2 we want.

For the sake of simplicity and comparison, a shortcut version of the n_2 simulation is also discussed, which we will call the theoretical simulation.

(III) Theoretical Simulation of n_2:

- Simulate the allocation of patients to the two treatment groups using the appropriate allocation procedure. After all the patients are being assigned, we get that there are n_E and n_C patients in the experimental and the control groups, respectively.
- Plug the simulated n_E, n_C, and the given parameters into the power formula (2) to calculate the theoretical power.
- Count the number of times for these calculated powers to be greater than or equal to 80% and divide it by 10,000. This is the probability of having a power of at least 80% for this sample size n.
- The simulation is repeated with different values of n until we find a minimum integer that generates at least a 90% confidence probability. This is the value of n_2, according to the theoretical simulation.

Comparing the results in Table 2 and Table 3 with the results in Table 1, we see that the simulation results decently resemble the theoretical results. Most of the simulation sample sizes are different from their respective theoretical sample sizes only by one or two, with the greatest discrepancies being four. Also, the theoretical simulation results (in parentheses) for n_2 are generally similar to the real simulation results, with the exceptions of CR, $\sigma_C = 4$ and GBC, $\sigma_C = 4$ with discrepancies of six and five, respectively.

Table 2. Simulated sample sizes of CR, UD, and GBC (10,000 replications, $\alpha = .05$, $\beta = .8$, $1 - \rho = .9$ and $\mu_E - \mu_C = 1$).

(σ_E, σ_C)	(1, 1)	(1, 2)	(1, 4)
n_1 (CR)	27	63	212
n_2 (CR)	29(28)*	72(70)	235(229)
n_1 (UD)	26	64	212
n_2 (UD)	26(27)	67(66)	222(222)
n_1 (GBC)	26	62	210
n_2 (GBC)	26(26)	65(65)	220(215)

* Numbers in () are produced using theoretical simulation

Table 3. Simulated sample sizes of SMLE and DBCD (10,000 replications, $\alpha = .05$, $\beta = .8$, $1 - \rho = .9$ and $\mu_E - \mu_C = 1$).

(σ_E, σ_C)	(1, 1)	(1, 2)	(1, 4)
n_1 (SMLE)	25	57	156
n_2 (SMLE)	27(29)	61(61)	162(160)
n_1 (DBCD, $\gamma = 1$)	25	56	155
n_2 (DBCD, $\gamma = 1$)	26(29)	60(60)	161(158)
n_1 (DBCD, $\gamma = 4$)	25	56	154
n_2 (DBCD, $\gamma = 4$)	26(29)	59(59)	161(158)

* Numbers in () are produced using theoretical simulation

5 DISCUSSION

The calculation of sample sizes for randomized designs has been demonstrated in the context of five specific randomized designs, including the Completely Randomized design (CR), Wei's (1977) Urn Design (UD), Generalized Biased Coin Designs (GBC) proposed by Smith (1984), Doubly-adaptive Biased Coin Design (DBCD) proposed by Eisele and Woodroofe (1995) and Hu and Zhang (2004), and lastly Sequential Maximum Likelihood Procedure (SMLE) proposed by Melfi and Page (2000). The sample sizes n_0, n_1, and n_2 were calculated using formulae developed and presented by Hu (2006), where n_0 is the sample size calculated by ignoring the randomness of allocation, n_1 is the sample size that produces a certain power on average when accounting for randomness, and n_2 is the sample size that produces a certain power with a certain probability when accounting for randomness. By comparing the theoretical results in Table 1 with our simulation results in Table 2 and Table 3, we could conclude that: although the simulation results are slightly different from their corresponding theoretical values in general, theoretical tables can still provide very good guidelines in predicting reality. Moreover, in practice, we could use the data in theoretical tables as initial values to do simulation, which will greatly reduce workload while preserving significant accuracy.

References

Eisele, J. and Woodroofe, M. (1995). Central limit theorems for doubly adaptive biased coin designs. *Ann. Statist.*, **23**, 234–254.

Hu, F. (2006). Sample size and power of randomized designs. *Unpublished manuscript.*

Hu, F. and Rosenberger, W.F. (2003). Optimality, variability, power: Evaluating response-adaptive randomization procedures for treatment comparisons. *Journal of the American Statistical Association*, 98, 671-678.

Hu, F. and Rosenberger, W.F. (2006). *The Theorey of Response-Adaptive Randomization in Clinical Trials.* John Wiley, Wiley Series in Probability and Statistics.

Hu, F. and Zhang L. X. (2004). Asymptotic properties of doubly adaptive biased coin designs for multi-treatment clinical trials. *Annals of Statistics.*, **32**, 268–301.

Melfi, V. and Page, C. (2000). Estimation after adaptive allocation. *Journal of Statistical Planning and Inference.*, 87 353-363.

Rosenberger, W. F. and Hu, F. (2004). Maximizing power and minimizing treatment failures in clinical trials. *Clinical Trials.*, **1**, 141–147.

Rosenberger, W. F. and Lachin, J.M. (2002). *Randomization in Clinical Trial.* Vol. John Wiley New York.

Smith, R. L. (1984). Properties of biased coin designs in sequential clinical trials. *Ann. Statist.*, **12**, 1018–1034.

Wei, L. J. (1977). A class of designs for sequential clinical trials. *J. Amer. Statist. Assoc.*, **72**, 382–386.