Chapter 8
Design and Analysis of Clinical Trial Simulations

Kazuhiko Kuribayashi

Abstract Clinical trial simulation is a powerful tool for supporting decision making in designing clinical trials, and plays an important role in clinical research and drug development. In clinical trial simulation, however, the design is often not well-considered and the results are empirically assessed. In this chapter, we present points to consider when planning a clinical trial simulation, and discuss how to design a clinical trial simulation employing a fractional factorial design and how to analyze the simulation results.

8.1 Introduction

Clinical trial simulation (CTS) is a process to mimic the conduct of a clinical trial on computers by generating the outcomes for each virtual patient based on the prespecified models and/or assumptions. CTS is a powerful tool for supporting decision making in designing clinical trials, and plays an important role in clinical research and drug development. The primary objective of CTS is to investigate the validity and robustness of study designs under various design scenarios and/or assumptions.

When planning clinical trials, complicated study designs such as adaptive designs are considered to achieve the objectives efficiently. Trial operating characteristics should be assessed at the planning stage of such complicated study designs. In particular, assessments of operating characteristics and factors that may influence them would help not only to select an optimal study design, but also to provide a guidance for trial monitoring. Since statistical theory for such study designs is often complicated and their operating characteristics are assessed analytically only under relatively strong assumptions, we usually rely on Monte Carlo simulations. CTS is relatively easily conducted to evaluate the operating characteristics under various practical settings. CTS is also useful for traditional fixed designs. In actual clinical

K. Kuribayashi (✉)
Pfizer Japan Inc., Tokyo, Japan
e-mail: kazuhiko.kuribayashi@pfizer.com

K. van Montfort et al. (eds.), *Developments in Statistical Evaluation of Clinical Trials*, DOI 10.1007/978-3-642-55345-5_8

trials, it is not unusual to deviate from the study protocol, and assessments of the effects of such deviations on the outcomes would be a key to study success.

In CTS, the number of simulations is often not objectively determined and the results are empirically assessed. Moreover, the design of factor arrangements is often not well-considered. It seems to be practical to perform simulations at all possible combinations of levels across all factors, which is a full factorial design. CTS generates virtual patient responses under a number of scenarios, which are combinations of levels of various factors. The number of combinations increases greatly with the increase in the number of factors and their levels. We often encounter difficulties to conduct simulations for all possible combinations of the levels with sufficient numbers of replications within a reasonable time. In such cases, if simulations are conducted with insufficient replications, then it is important to evaluate the Monte Carlo error. On the other hand, we can reduce the number of combinations of levels of factors by employing a fractional factorial design, which is a factorial design in which only an adequately chosen subset of the combinations required for the full factorial design is selected to be run (e.g., [6]).

In this chapter, we present points to consider when planning CTS and discuss how to design CTS and how to analyze the results. In Sect. 8.2, protocol development of CTS and how to determine the number of simulations based on the Monte Carlo error are described. In Sect. 8.3, the design of CTS using orthogonal array and the analysis of simulation results are presented. An example of an adaptive group sequential design is illustrated in Sect. 8.4. Finally, some remarks are provided in Sect. 8.5.

8.2 Planning of Clinical Trial Simulations

8.2.1 Protocol Preparation

As poorly designed and poorly conducted clinical trials produce questionable results, poorly designed and poorly conducted CTS also make inappropriate choices of study designs and statistical methods. Hence, CTS should be planned with similar rigor as clinical trials, in particular, if the purpose of CTS is to provide information on decision making in designing clinical trials. Planning the CTS, "protocol", which describes what the objectives of the simulation are, how the simulation is to be performed and how the results are assessed, should be prepared as clinical trials [2, 5, 12]. The protocol also includes the rationale for all the specifications of the CTS plan. An example of the contents of the protocol is as follows.

Objectives of the Simulation Study Clearly defined objectives of the simulation study should be stated in the protocol. This includes how to assess questions of interest by simulation and how to leverage the simulation results to decision making.

Scenarios and Factors to Investigate with Rationale Scenarios of the clinical outcome to be investigated by simulation should be described along with some rationale. The scenarios include favorable, unfavorable and highly possible ones. Factors and their levels to be examined should be also described.

Simulation Study Design CTS usually generates virtual patient responses under combinations of levels of various factors. This is considered as a factorial experiment. The design of factor arrangements should be well-considered. The factor arrangements in the simulation, such as full factorial design, fractional factorial design or split-plot design (e.g., [6]), should be explained.

Data Generation Method A thorough description of data generation methods should be provided. This includes the rationale for selections of assumed distributions, required parameters for statistical models and correlation structure of the covariates.

The random number generation method should be described. The quality of simulation depends very much on the quality of the pseudorandom numbers. Unreliable algorithms should not be employed.

The data generated should simulate situations that enable to generalize the simulation results, and should be checked by using some statistics, such as summary statistics for distributions of the covariates and Kaplan-Meier estimates for time-to-event data.

It might be useful to simulate data by bootstrapping or permutation from real clinical trial data for creating resemblance to reality.

It is also useful to apply the inclusion and exclusion criteria of the clinical trial to generated data.

Assessments The operating characteristics quantifying the performance of the study design, such as power, expected sample size and so on, to be evaluated in CTS, should be defined.

Determination of the Number of Simulation Replications The rationale for the number of simulation replications should be stated. The number of simulations can be determined based on the Monte Calro error. Details are described in the next section.

Statistical Evaluation The analysis methods for the simulation results should be stated. How to handle ill-conditioned cases, such as failure to estimate parameters of interest due to non-convergence and/or infrequent events, should be described.

8.2.2 Determination of the Number of Simulation Replications

The estimated accuracy of operating characteristics, which is the amount of the Monte Carlo error, depends on the number of simulation replications R. Once the target amount of the Monte Carlo error is chosen, the number of simulation

replications is determined using the inversely proportional relationship between the Monte Carlo error and the square root of the number of replications [8].

Let θ be an operating characteristic to be evaluated by simulation, and $\hat{\theta}^{(R)}$ the estimate based on the R simulations. For instance, when the operating characteristic to be evaluated by simulation is the power or the probability of type I error, letting $I[\cdot]$ be an indicator function which equals 1 when the argument is true, 0 otherwise, $z^{(r)}$ the test statistics at the rth simulation and c the critical value, the estimate of the power or the probability of type I error is provided by

$$\hat{\theta}^{(R)}_{\text{power}} = \frac{1}{R}\sum_{r=1}^{R} I[z^{(r)} > c] \, .$$

The estimate of the expected sample size based on the R simulation replications is provided by

$$\hat{\theta}^{(R)}_{N} = \frac{1}{R}\sum_{r=1}^{R} N^{(r)} \, ,$$

where $N^{(r)}$ denotes the sample size at the rth simulation. The variability of the estimated operating characteristics is quantified by the Monte Carlo error

$$\text{MCE}(\hat{\theta}^{(R)}) = \sqrt{V(\hat{\theta}^{(R)})} \, ,$$

where $V(\cdot)$ denotes the variance [8]. To estimate the Monte Carlo error, the R simulation replications need to be replicated a sufficient number of times. This would be impractical since an additional investment of time is required. If $\hat{\theta}^{(R)}$ is asymptotically normal, then the estimated Monte Carlo error is obtained as

$$\widehat{\text{MCE}}_{\text{asym}}(\hat{\theta}^{(R)}) = \frac{\hat{\sigma}_{\theta}}{\sqrt{R}} = \frac{1}{\sqrt{R}}\sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left(S^{(r)} - \frac{1}{R}\sum_{r=1}^{R} S^{(r)}\right)^2} \, , \tag{8.1}$$

where $S^{(r)}$ denotes an outcome related to the operating characteristic at the rth simulation, such as $S^{(r)} = I[z^{(r)} > c]$ for the power or the probability of type I error and $S^{(r)} = N^{(r)}$ for the expected sample size. If $\hat{\theta}^{(R)}$ is not asymptotically normal, the bootstrap method can be employed. B sets of bootstrap samples with size R, $\boldsymbol{S}^*_1, \boldsymbol{S}^*_2, \ldots, \boldsymbol{S}^*_B$, are drawn with replacement from $\boldsymbol{S} = \{S^{(1)}, S^{(2)}, \ldots, S^{(R)}\}$ generated by R simulations, and $\hat{\theta}^{(R)}(\boldsymbol{S}^*_1)$, $\hat{\theta}^{(R)}(\boldsymbol{S}^*_2)$,..., $\hat{\theta}^{(R)}(\boldsymbol{S}^*_B)$ are calculated for each bootstrap sample. A bootstrap estimate of the Monte Carlo error is provided by

$$\widehat{\text{MCE}}_{\text{boot}}(\hat{\theta}^{(R)}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^{(R)}(\boldsymbol{S}^*_b) - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(R)}(\boldsymbol{S}^*_b)\right)^2} \, .$$

The number of simulation replications R can be determined by the target amount of the Monte Carlo error and the variation between simulations σ_θ in (8.1). When an operating characteristic to be evaluated is the binomial proportion, such as the power or the probability of type I error, the variation between simulations is obtained as

$$\sigma_\theta = \sqrt{Q(1-Q)}\,,$$

where Q denotes the assumed value of the proportion. Letting MCE' be a target amount of the Monte Carlo error, the required number of simulations is

$$R' = \left(\frac{\sigma_\theta}{\mathrm{MCE}'}\right)^2 . \tag{8.2}$$

For example, when estimating the probability of type I error with the Monte Carlo error 0.001 in a one-sided test with significance level 0.025, 24,375 simulations are required. In the case of Monte Carlo error 0.005, 975 simulations are required. It is not unusual to have much uncertainty in the assumed value of the power. In such case, the calculation using $Q = 0.5$, which gives the largest variation between simulations, is on the safe side. When $Q = 0.5$, 10,000 simulations are required to achieve a 0.005 for the Monte Carlo error. This means that the Monte Carlo error of the binomial probability estimated by 10,000 simulations is at most 0.005.

When the variation between simulations is unknown, such as the expected sample size, it can be estimated by simulation. First, R simulations are tentatively conducted and $\{S^{(1)}, S^{(2)}, \ldots, S^{(R)}\}$ are obtained. Next, $R_1, R_2, \ldots, R_p$ samples are drawn with replacement. That is, $\boldsymbol{S}_1^* = \{S^{(1)}, \ldots, S^{(R_1)}\}$, $\boldsymbol{S}_2^* = \{S^{(1)}, \ldots, S^{(R_2)}\}$, …, $\boldsymbol{S}_p^* = \{S^{(1)}, \ldots, S^{(R_p)}\}$ are generated. The Monte Carlo error is estimated in each set, and the variation between simulations σ_θ is estimated as the slope by applying the least-squares method to the paired data, $\left(\frac{1}{\sqrt{R_1}}, \widehat{\mathrm{MCE}}_{\mathrm{boot}}(\hat{\theta}^{(R_1)})\right)$, $\left(\frac{1}{\sqrt{R_2}}, \widehat{\mathrm{MCE}}_{\mathrm{boot}}(\hat{\theta}^{(R_2)})\right)$, …, $\left(\frac{1}{\sqrt{R_p}}, \widehat{\mathrm{MCE}}_{\mathrm{boot}}(\hat{\theta}^{(R_p)})\right)$.

8.2.3 *Determination of the Number of Bootstrap Samples*

When employing the bootstrap method to estimate the Monte Carlo error, the accuracy depends on the number of bootstrap samples B. The number of bootstrap samples is determined so that the probability that the relative error of the bootstrap estimates of the Monte Carlo error falls within a certain range is ensured [1]. That is, we choose B such that

$$1 - \omega = \Pr\left(1 - \gamma < \frac{\widehat{\mathrm{MCE}}_{\mathrm{boot}}(\hat{\theta}^{(R)})^2}{\hat{V}(\hat{\theta}^{(R)})} < 1 + \gamma\right),$$

where ω and γ denote a small probability and a small positive value, respectively. Suppose that the distribution of $\hat{\theta}^{(R)}$ is approximately normal and $\chi^2_{B-1} = (B-1)\,\widehat{\mathrm{MCE}}_{\mathrm{boot}}(\hat{\theta}^{(R)})^2 \,/\, \hat{V}(\hat{\theta}^{(R)})$ has approximately a chi-squared distribution with $(B-1)$ degrees of freedom. Since B is large enough to ignore the difference between B and $B-1$, the approximation

$$\begin{aligned}
&\Pr\left(1-\gamma < \frac{\widehat{\mathrm{MCE}}_{\mathrm{boot}}(\hat{\theta}^{(R)})^2}{\hat{V}(\hat{\theta}^{(R)})} < 1+\gamma\right) \\
&\quad\approx \Pr\left(B(1-\gamma) < \chi^2_B < B(1+\gamma)\right) \\
&\quad\approx \Pr\left(B(1-\gamma) < B + \sqrt{2B}\frac{\hat{\theta}^{(R)}}{\sqrt{\hat{V}(\hat{\theta}^{(R)})}} < B(1+\gamma)\right) \\
&\quad= 1 - 2\Phi\left(-\sqrt{\frac{B}{2}}\gamma\right)
\end{aligned}$$

can be obtained. The number of bootstrap samples to achieve a relative error less than γ with probability $1-\omega$ is approximately

$$B \approx \frac{2\left(\Phi^{-1}\left(\frac{\omega}{2}\right)\right)^2}{\gamma^2}\,.$$

For example, 769 bootstrap samples are required to achieve a relative error ranged from 0.9 to 1.1 with probability 0.95.

8.3 Design and Analysis of CTS by Orthogonal Arrays

In CTS, operating characteristics quantifying the performance of the study design are evaluated under a number of scenarios, which are combinations of the levels across various factors. This is considered as a factorial experiment. In practice, CTS is often conducted at all possible combinations of levels across all factors, which is a full factorial experiment. In that case, the number of combinations increases greatly with an increase in the number of factors and their levels. For example, with ten factors each taking two levels, a full factorial experiment would have $2^{10} = 1{,}024$ combinations in total. This means R simulations at each combination have to be replicated 1,024 times. It can easily be imagined that such a full factorial experiment with a sufficient number of simulations for each combination requires a great deal of time. In that case, it might be difficult to perform CTS with sufficient replications. On the other hand, we can try to reduce of the number of combinations of the levels of the factors.

Table 8.1 Orthogonal array for 2 levels, $L_8(2^7)$

	Columns						
Run	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2
	a	b	a b	c	a c	b c	a b c

A full factorial experiment evaluates the main effect of each factor as well as the effects of interactions between factors. For ten factors, each taking two levels, the full factorial experiment requires 1,024 simulation runs and allows to evaluate 1,013 interactions including ${}_{10}C_2 = 45$ two-factor interactions, ${}_{10}C_3 = 120$ three-factor interactions, . . ., ${}_{10}C_{10} = 1$ ten-factor interactions. However, usually it is very difficult to interpret higher-order interactions, such as more than three factors. Such higher-order interactions could be negligible. If so, there is no need to employ a full factorial experiment. Rather a fractional factorial experiment, which is a factorial experiment in which only an adequately chosen subset of the combinations required for the full factorial experiment is selected to be run, may be useful and the factors are easily assigned by Taguchi's orthogonal array (e.g., [6]).

Table 8.1 shows an example of an orthogonal array for 2 levels. This table is represented by $L_8(2^7)$, where "L" stands for Latin squares because orthogonal array is an expansion of Latin squares, "8" indicates the number of rows, "2" means the number of levels and "7" is the number of columns. When selecting any two columns from this table, they include four types of combinations, (1,1), (1,2), (2,1) and (2,2), with the same frequency. We allocate a factor to one of the columns and assign 1 for one level and 2 for the other level, and then conduct simulations for eight combinations of the levels of the factors.

When the number of factors is three, this is equivalent to the full factorial experiment. But if some interactions are negligible, then we can allocate more than three factors. Consider a simulation study with four factors, say A, B, C and D, each taking two levels, and no interactions between the factors. The full factorial experiment requires 16 simulation runs. In contrast, a fractional factorial design using the orthogonal array presented in Table 8.2 requires 8 simulation runs. The four factors, A, B, C and D are allocated to 4 columns out of 7 and 8 combinations of the levels of the factors are determined. We can examine the main effects of the factors based on results of the 8 simulation runs.

Table 8.2 Assignment of factors in an orthogonal array

Run	Columns 1	2	3	4	5	6	7	Combinations
1	1	1	1	1	1	1	1	$A_1B_1C_1D_1$
2	1	1	1	2	2	2	2	$A_1B_1C_1D_2$
3	1	2	2	1	1	2	2	$A_1B_2C_2D_2$
4	1	2	2	2	2	1	1	$A_1B_2C_2D_1$
5	2	1	2	1	2	1	2	$A_2B_1C_2D_2$
6	2	1	2	2	1	2	1	$A_2B_1C_2D_1$
7	2	2	1	1	2	2	1	$A_2B_2C_1D_1$
8	2	2	1	2	1	1	2	$A_2B_2C_1D_2$
	A	B	C				D	

In the example above, we used an array with 2 levels in each factor and 7 columns for simplicity. If each factor takes the same number of levels, then corresponding orthogonal arrays are available. For factors with three levels, $L_{27}(3^{13})$, which has 27 rows and 13 columns, is available. Orthogonal arrays can handle factors taking different number of levels. For example, when allocating a factor taking 4 levels to $L_8(2^7)$, we choose any two columns and allocate the 4 levels to each of 4 types of combinations, (1,1), (1,2), (2,1) and (2,2).

Simulation results based on the orthogonal array can be analyzed as a factorial experiment since all the factors are orthogonal. In the case of Table 8.2, the total sum of squares S_T is the summation of the sum of squares of the factors, A, B, C, D and the error:

$$S_T = S_A + S_B + S_C + S_D + S_e \ ,$$

where S_e denotes the sum of squares of the error, and the effects of the factors are evaluated by analysis of variance.

8.4 An Illustrative Example: Adaptive Group Sequential Trial

We describe a process of CTS using an example, that applies an adaptive group sequential trial.

Consider a one-sided test with significance level $\alpha(= 0.025)$ of the null hypothesis $H_0 : \mu_x = \mu_y$ against the alternative hypothesis $H_1 : \mu_x > \mu_y$ in a confirmatory trial with two treatments. Now suppose the response of the test treatment $x \sim N(\mu_x, \sigma^2)$, that of the control $y \sim N(\mu_y, \sigma^2)$, and $\delta = (\mu_x - \mu_y)/\sigma$.

This trial employs a group sequential design with the sample size $2n_0$ allowing an interim analysis with $2tn_0$ $(0 < t < 1)$ subjects. When the test statistic doesn't cross the boundary at the interim analysis, the sample size is re-estimated based on

the conditional power and increased up to $2rn_0$ ($r > 1$). Let $2n$ be the re-estimated sample size, $\bar{x}_1$ and $\bar{y}_1$ denote the sample means at the interim analysis in each treatment group, respectively, and $\hat{\delta}_1 = (\bar{x}_1 - \bar{y}_1)/\sigma$. The test statistic at the interim analysis is given by

$$z_1 = \frac{\hat{\delta}_1}{\sqrt{\frac{2}{tn_0}}}.$$

At the final analysis, the weighted Wald statistic

$$z = z_1\sqrt{t} + z_2\sqrt{1-t}$$

is used as the test statistic [3], where

$$z_2 = \frac{\hat{\delta}_2}{\sqrt{\frac{2}{n-tn_0}}} = \frac{\bar{x}_2 - \bar{y}_2}{\sigma\sqrt{\frac{2}{n-tn_0}}}$$

denotes the Wald statistic based on $2(n - tn_0)$ subjects entered after the interim analysis. Cui et al. [3] showed that the weighted Wald statistic z has the same distribution as with the original sample size $2n_0$ under the null hypothesis. So we can use the original boundary without inflation of the probability of type I error even when increasing the sample size.

The conditional power given z_1 at the interim analysis is provided by

$$CP_{\hat{\delta}_1} = \Pr(z > c \mid z_1) = 1 - \Phi\left(c\sqrt{\frac{1}{1-t}} - z_1\sqrt{\frac{t}{1-t}} - \frac{\hat{\delta}_1}{\sqrt{\frac{2}{n-tn_0}}}\right),$$

where c denotes the boundary at the final analysis and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. The sample size to achieve the conditional power CP is obtained as

$$n = \frac{2\left(c\sqrt{\frac{1}{1-t}} - z_1\sqrt{\frac{t}{1-t}} - u_{1-CP}\right)^2}{\hat{\delta}^2} + tn_0,$$

where $u_{1-CP} = \Phi^{-1}(1 - CP)$. The boundaries for efficacy and futility stopping are calculated based on O'Brien-Fleming type α-spending function [9].

Suppose that we would like to assess the influence of the minimum requirement for sample size increase (A), target conditional power (B), upper limit of sample size (C) and timing of interim analysis (D) on the overall power, and to estimate the optimal combination of the levels of the factors, and also evaluate the expected

Table 8.3 Factors and their levels

Factor	Levels
Minimum requirement for sample size increase (A)	$z_1 >$ lower boundary, $CP_{\hat{\delta}_1} > 0.5$
Target conditional power (B)	$CP = 0.8, CP = 0.9$
Upper limit of sample size (C)	$r = 2, r = 3$
Timing of interim analysis (D)	$t = 0.3, t = 0.5$

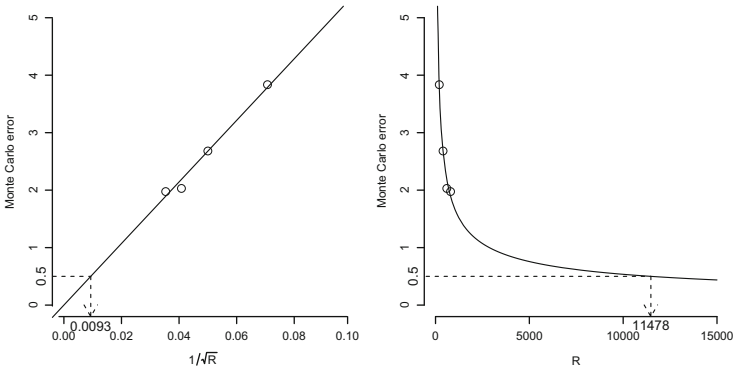

Fig. 8.1 Plots of the four pair of the estimated Monte Carlo error and the size of bootstrap samples and the line fitted by the least-squares method

sample size at that combination. Table 8.3 shows the levels of interest of the factors. In addition, we have interests in 2 two-factor interactions, $A \times B$ and $B \times C$, while the others are negligible.

The number of simulations is determined based on the Monte Carlo error in estimating the power and the expected sample size. For the power, 10,000 simulations are required to estimate it with 0.005 of Monte Carlo error when the variation between simulations $\sigma_\theta = 0.5$, which is largest. The variation between simulations for the expected sample size is estimated by simulation. Thousand simulations are conducted using the following levels of the factors shown in Table 8.3: $A : z_1 >$ Lower boundary, $B : CP = 0.9$, $C : r = 2$, $D : t = 0.5$. From the simulation results $\{S^{(1)}, \ldots, S^{(1,000)}\}$, four sets of bootstrap samples with the size $\{R_1, R_2, R_3, R_4\} = \{200, 400, 600, 800\}$ are drawn with replacement, and the Monte Carlo error for each bootstrap sample is calculated. The variation between simulations is estimated as $\hat{\sigma}_\theta = 53.57$ by the least-squares method applied to the four pairs of the estimated Monte Carlo error and the size of the bootstrap samples. Figure 8.1 shows the plots of the four pair values and the fitted line. The number of simulations required to estimate the expected sample size with 0.5 of the Monte Carlo error is calculated by assigning $\hat{\sigma}_\theta = 53.57$ and $\mathrm{MCE}' = 0.5$ to (8.2). This provides $R' = 11{,}478$. Taking into consideration the above calculations, we determine to conduct 10,000 simulations.

Table 8.4 Assignment of factors and simulation results

	Columns								Simulation results[a]	
Run	1	2	3	4	5	6	7	Combinations	Power	ESS
1	1	1	1	1	1	1	1	$A_1B_1C_1D_1$	0.9508	167.97
2	1	1	1	2	2	2	2	$A_1B_1C_2D_2$	0.9422	218.28
3	1	2	2	1	1	2	2	$A_1B_2C_1D_2$	0.9156	149.38
4	1	2	2	2	2	1	1	$A_1B_2C_2D_1$	0.9840	251.75
5	2	1	2	1	2	1	2	$A_2B_1C_1D_2$	0.9197	136.25
6	2	1	2	2	1	2	1	$A_2B_1C_2D_1$	0.9877	216.65
7	2	2	1	1	2	2	1	$A_2B_2C_1D_1$	0.9525	147.34
8	2	2	1	2	1	1	2	$A_2B_2C_2D_2$	0.9425	200.95
	A	B	$A \times B$	C		$B \times C$	D			

[a] Based on 10,000 replications

Table 8.5 Analysis of variance for the simulation result

Factors	Df	Sum Sq	Mean Sq	F value	Pr($>F$)	Prop SS
A	1	0.00001200	0.00001200	29.6420	0.115641	0.002501
B	1	0.00000421	0.00000421	10.3827	0.191572	0.000876
C	1	0.00173460	0.00173460	4282.9753	0.009727	0.361407
D	1	0.00300313	0.00300313	7415.1235	0.007393	0.625704
$A \times B$	1	0.00004512	0.00004512	111.4198	0.060132	0.009402
$B \times C$	1	0.00000012	0.00000012	0.3086	0.677171	0.000026
Residuals	1	0.00000040	0.00000040			0.000084
Total	7	0.00479960				

Table 8.6 The point estimates and 95 % confidence intervals of means at each combination of the upper limit of the sample size (C) and the timing of the interim analysis (D)

		Estimate	95 % C.I.	
C_1	$r = 2$	0.9347	0.9335	0.9358
C_2	$r = 3$	0.9641	0.9630	0.9652
D_1	$t = 0.3$	0.9688	0.9676	0.9699
D_2	$t = 0.5$	0.9300	0.9289	0.9311

The allocation of the factors and the simulation results are shown in Table 8.4 and the analysis of variance (ANOVA) table is shown in Table 8.5. This indicates that the upper limit of the sample size (C) and timing of the interim analysis (D) have some effect on the power.

The point estimates and 95 % confidence intervals of means at each combination of the upper limit of the sample size (C) and the timing of the interim analysis (D) are calculated based on the fitted ANOVA model, and shown in Table 8.6. This table

suggests that the combination of C_2 ($r = 3$) and D_1 ($t = 0.3$) is optimal. The point estimates and 95 % confidence intervals of means at the optimal combination based on the fitted ANOVA model are 0.9835 for the power with 95 % confidence interval (0.9821, 0.9848) and 231.76 for the expected sample size with 95 % confidence interval (220.23, 243.30).

This example was implemented by R [11].

8.5 Concluding Remarks

Clinical trial simulations are a statistical experiment, and should be appropriately performed with careful planning. Even if advanced methodologies/technologies are employed, incomplete inputs produce incomplete outputs or, as it is often said, "garbage in, garbage out." CTS should be planned with similar rigor as clinical trials, and conducted with the following two points in mind:

1. To achieve the given purpose of the simulation study, what is the best way to obtain appropriate information with the smallest number of simulations in total?
2. To draw the accurate conclusion, how should the simulation results including the Monte Carlo error be analyzed?

In reporting clinical trials, standard errors and 95 % confidence intervals are routinely presented with point estimates. In reporting CTS, only point estimates are presented in practice. As a guidance for reporting simulation studies for statistical methods, it is pointed out that all reporting should make it easy for the reader to assess the quality of the experimental work and the accuracy of the results [7]. In the same way, reporting CTS should routinely include the Monte Carlo error and 95 % confidence intervals. The 95 % confidence interval is given by

$$\left(\hat{\theta}^{(R)} - 1.96\,\widehat{\mathrm{MCE}}_{\mathrm{asym}}(\hat{\theta}^{(R)}),\quad \hat{\theta}^{(R)} + 1.96\,\widehat{\mathrm{MCE}}_{\mathrm{asym}}(\hat{\theta}^{(R)})\right),$$

where $\hat{\theta}^{(R)}$ is asymptotically normal. If it is not normal, but the distribution is symmetric about $\hat{\theta}^{(R)}$, the 95 % confidence interval is estimated by the 2.5 and 97.5 percentile of $\hat{\theta}^{(R)}(\boldsymbol{S}_1^*), \hat{\theta}^{(R)}(\boldsymbol{S}_2^*), \ldots, \hat{\theta}^{(R)}(\boldsymbol{S}_B^*)$,

$$\left(\hat{\theta}^{(R)}_{B[0.025]},\quad \hat{\theta}^{(R)}_{B[0.975]}\right).$$

If it is not symmetric, the interval is given by

$$\left(2\hat{\theta}^{(R)} - \hat{\theta}^{(R)}_{B[0.975]},\quad 2\hat{\theta}^{(R)} - \hat{\theta}^{(R)}_{B[0.025]}\right)$$

(e.g., [4]). In addition, the limitations of the conclusion and recommendation from the simulation study should be addressed in the reporting of CTS.

In this chapter, we discussed CTS with factors, which each takes fixed level values. Taking into account uncertainties, including randomness in the sampling of subjects, uncertainty about the baseline characteristics of the subject population and uncertainty about the treatment's clinical effects, we can consider Bayesian CTS, which simulates parameter values from probability distributions that represent the current state of knowledge about the parameters [10]. Bayesian CTS accounts for all sources of uncertainty and allows more realistic assessments of the outcomes of individual clinical trials and sequences of clinical trials for the purpose of decision making. In Bayesian CTS as well, the concept of the experimental design discussed here is important. This concept is applicable not only to CTS, but also to assessment of statistical methodologies.

References

1. Booth, J.G., Sarkar, S.: Monte Carlo approximation of bootstrap variances. *The American Statistician* **52**, 354–357 (1998)
2. Burton, A., Altman, D.G., Royston, P., Holder, R.L.: The design of simulation studies in medical statistics. *Statistics in Medicine* **25**, 4279–4292 (2006)
3. Cui, L., Hung, H.M.J., Wang, S.J.: Modification of sample size in group sequential clinical trials. *Biometrics* **55**, 853–857 (1999)
4. Davison, A.C., Hinkley, D.V.: *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge, UK (1997)
5. Gaydos, B., Anderson, K.M., Berry, D., Burnham, N., Chuang-Stein, C., Dudinak, J., Fardipour, P., Gallo, P., Givens, S., Lewis, R., Maca, J., Pinheiro, J., Pritchett, Y., Krams, M.: Good practices for adaptive clinical trials in pharmaceutical product development. *Drug Information Journal* **43**, 539–556 (2009)
6. Giesbrecht, F.G., Gumpertz, M.L.: *Planning, Construction, and Statistical Analysis of Comparative Experiments*. John Wiley & Sons: Hoboken, NJ (2004)
7. Hoaglin, D.C., Andrews, D.F.: The reporting of computation-based results in statistics. *The American Statistician* **29**, 122–126 (1975)
8. Koehler, E., Brown, E., Haneuse, S.J.P.A.: On the assessment of Monte Carlo error in simulation-based statistical analysis. *The American Statistician* **63**, 155–162 (2009)
9. Lan, K.K.G., DeMets, D.L.: Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–663 (1983)
10. O'Hagan, A., Stevens, J.W., Campbell, M.J.: Assurance in clinical trial design. *Pharmaceutical Statistics* **4**, 187–201 (2005)
11. R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2012). URL http://www.R-project.org/
12. Smith, M.K., Marshall, A.: Importance of protocols for simulation studies in clinical drug development. *Statistical Methods in Medical Research* **20**, 613–622 (2011)