An efficient procedure for calculating sample size through statistical simulations

Kazushi Maruo
Department of Clinical Epidemiology, Translational Medical Center,
National Center of Neurology and Psychiatry
Keisuke Tada
Biostatistics & Programming, Sanofi K. K.
Ryota Ishii
Clinical Data Science Dept., Kowa Company, LTD.
and
Masahiko Gosho
Department of Clinical Trial and Clinical Epidemiology,
Faculty of Medicine, University of Tsukuba

May 30, 2017

Abstract

While planning clinical trials, when simple formulae are unavailable to calculate sample size, statistical simulations are used instead. However, one has to spend much computation time obtaining adequately precise and accurate simulated sample size estimates, especially when there are many scenarios for the planning and/or the specified statistical method is complicated. In this paper, we summarize the theoretical aspect of statistical simulation-based sample size calculation. Then, we propose a new simulation procedure for sample size calculation by fitting the probit model to simulation result data. From the theoretical and simulation-based evaluations, it is suggested that the proposed simulation procedure provide more efficient and accurate sample size estimates than ordinary algorithm-based simulation procedure especially when estimated sample sizes are moderate to large, therefore it would help to dramatically reduce the computational time required to conduct clinical trial simulations.

Keywords: clinical trial design; finite sample bias; probit model

1 Introduction

Sample size calculation plays an important role in estimating the costs and the success probability (e.g., power for statistical tests) of clinical trials. When the design of a planned trial is simple and the targeted endpoint is assumed to follow an ordinary theoretical distribution (e.g., normal, binomial, or exponential), the sample size is usually calculated using a formula (e.g., see Julious (2004); Julious and Campbell (2012); Schoenfeld (1983)). Additionally sample size formulae for complicated distributions and/or designs have been developed recently (e.g., see Zhang and Pulkstenis (2016); Zhu (2017)), such formulae are not always available. For example, adaptive designs such as sample size re-estimation (Cui et al., 1999; Chen et al., 2004) and seamless phase II/III design (Bretz et al., 2006; Maca et al., 2006) have been implemented in the drug development process. Also, Bayesian analyses with Markov chain Monte Carlo methods are sometimes conducted as the primary analysis of a clinical trial, especially in the early phase of a drug development process (Thall et al., 2003; Tighiouart et al., 2005). More recently, statistical tests for treatment effects based on the multiple imputation methods (Rubin, 1987) for missing values have been often applied as primary statistical analyses. In these complicated cases, sample size formulae are usually unavailable and, therefore, the sample sizes need to be calculated using statistical simulations (e.g., see Chow and Chang (2011)).

The following simple simulation procedure is commonly used to calculate sample sizes. Let the simulated power for a specified sample size and specified settings (e.g., statistical method, success criteria, effect size, etc.) be the success proportion in the number of simulations for the binary simulation result data (success, failure). The calculation of the simulated power for a specified setting starts from a sufficiently small sample size, which is then gradually increased until the simulated power exceeds a specified level, for instance, 0.8 or 0.9. This is an ordinary approach to estimate the sample size using statistical simulations. The estimated required sample size is the sample size at the end of the simulations. In order to ensure the precision of the estimated sample sizes using this process, many simulations need to be conducted. In addition, multiple simulation factors (mean difference, SD, baseline hazard, etc.) with multiple levels are usually considered in clinical trial simulations. Thus, a statistical simulation-based sample size calculation

may be a time-consuming process and, therefore, it accounts for a considerable part of the cost of the design process of a clinical trial. Furthermore, although some researchers have focused on the general theory of a statistical simulation in medical research (e.g., see Burton et al. (2006)), no studies have focused on the theoretical aspect of using statistical simulations to calculate a sample size.

In this paper, we provide the inference theory for a estimated sample size for a clinical trial based on the ordinary statistical simulation procedure. Then we propose a procedure that applies the probit regression model to the binary simulation result data. We show that the accuracy and precision of the estimated sample size improve by estimating sample sizes with the statistical model instead of the ordinary algorithm-based method. We also demonstrate that the proposed simulation procedure reduces the computational time of clinical trial simulations dramatically. In Section 2, we present the theoretical aspect of a sample size estimation based on the ordinary statistical simulation procedure, and then provide a new viewpoint that the probit model is applicable to statistical simulation results. In Section 3, we describe the effect of a model misspecification in the probit model and provide a new simulation procedure. Then, in Section 4, we compare the estimation performances between the ordinary and the proposed procedures, both theoretically and through simulations. We summarize our results in Section 5.

2 Simulation-based sample size calculation: theoretical aspect

In this section, we consider a simulation approach in which a sample size is calculated based on statistical simulations, such that some successful probability for the specified statistical decision tool (e.g., power for a two-sample t-test) achieves a specified level, p (e.g., 0.8 or 0.9). For a given sample size, n, and a given number of simulations, m, the number of successes, X_n, follows a binomial distribution, $Bin(m, \pi_n)$, where π_n is the true success probability for n. The targeted sample size, n_p, is defined as the smallest sample size such that π_n exceeds p. In the following discussion, statistical hypotheses and confidence intervals are both-tailed.

2.1 Ordinary simulation procedure

In this section, we provide the theoretical results of a commonly used simulation procedure for sample size estimation. The maximum likelihood estimator (MLE) for the binomial proportion π_n and the variance of the MLE are given as $\hat{\pi}_{n(m)} = x_n/m$ and $Var(\hat{\pi}_{n(m)}) = \pi_n(1-\pi_n)/m$, respectively. Let statistical simulations start from a sufficiently small sample size, which is then successively increased by one. For each sample size, $\hat{\pi}_{n(m)}$ is calculated. When m is sufficiently large, $Pr(\hat{\pi}_{n(m)} < p)$ can be approximated by

$$Pr(\hat{\pi}_{n(m)} < p) \simeq \Phi\left\{\frac{\sqrt{m}(p-\pi_n)}{\sqrt{\pi_n(1-\pi_n)}}\right\},$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The estimator of the sample size, $\hat{n}_{p(m)}$, is usually defined as the smallest sample size such that $\hat{\pi}_{n(m)}$ exceeds p in the simulations. Thus, the probability function of $\hat{n}_{p(m)}$ is given by

$$f_{p(m)}(n) = Pr(\hat{n}_{p(m)} = n) = \prod_{i=1}^{n-1} Pr(\hat{\pi}_{i(m)} < p)\{1 - Pr(\hat{\pi}_{n(m)} < p)\}.$$

The expectation and variance of $\hat{n}_{p(m)}$ are given by

$$E(\hat{n}_{p(m)}) = \sum_{n=1}^{\infty} n f_{p(m)}(n) \text{ and } Var(\hat{n}_{p(m)}) = \sum_{n=1}^{\infty} n^2 f_{p(m)}(n) - E(\hat{n}_{p(m)})^2, \quad (1)$$

respectively. In practice, the sum is calculated to a sufficiently large number. In this paper, for the above moments, we calculate the sum to $4n_p$ because, in general, π_{4n_p} is very close to 1 for $p \geq 0.8$. Unless m is considerably large, $\hat{n}_{p(m)}$ has the bias, $E(\hat{n}_{p(m)}) - n_p$, because of the simulation procedure used to select the smallest sample size. In the following, we refer to this procedure as the ordinary procedure.

If $\pi_{n_p} = p$ exactly, we have $\lim_{m\to\infty} E(\hat{n}_{p(m)}) = n_p + 1/2$ and $\lim_{m\to\infty} Var(\hat{n}_{p(m)}) = 1/4$ because $\lim_{m\to\infty} Pr(\hat{n}_{p(m)} = n_p) = \lim_{m\to\infty} Pr(\hat{n}_{p(m)} = n_p+1) = 1/2$. In practice, however, π_{n_p} is larger than p (e.g., $\pi_{n_p} = 0.802$ for $p = 0.8$), and we have $\lim_{m\to\infty} E(\hat{n}_{p(m)}) = n_p$ and $\lim_{m\to\infty} Var(\hat{n}_{p(m)}) = 0$.

2.2 Applying a probit model to simulation result data

Usually, sample size formulae for statistical tests based on the normal distribution or the normal approximation are described as

$$n_p = \left\lceil \delta^{-2} \left\{ z_{1-\alpha/2} + z_p \right\}^2 \right\rceil, \tag{2}$$

where δ is equal to the product of a constant and the so-called "effect size," $z_p = \Phi^{-1}(p)$ is the $100p$ percentile of the standard normal distribution, $z_{1-\alpha/2}$ is the critical value of the statistical test at a significance level of α, and $\lceil \cdot \rceil$ is a ceiling function. For example, δ =\{mean difference between two groups\}/[$\sqrt{2}$\{SD for each group\}] for a two-sample t-test, and δ =log\{hazard ratio\}/2 for a log rank test (Schoenfeld, 1983). However, as mentioned earlier, more complicated settings appear in practical clinical trials where sample size formulae are not, in general, available.

In this study, we focus on situations in which no sample size formula is available and δ is unknown, and then develop the simulation method using a probit function to estimate the sample size. Here, we consider situations where equation (2) holds for unknown δ. When equation (2) holds approximately, π_n can be described as the following probit model structure by solving the equation, $n = \delta^{-2} \left\{ z_{1-\alpha/2} + \Phi^{-1}(\pi_n) \right\}^2$, for π_n:

$$\pi_n = \Phi(\beta_0 + \beta_1 \sqrt{n}), \tag{3}$$

where $\beta_0 = -z_{1-\alpha/2}$, and $\beta_1 = \delta$. Let $Y_{n(j)}$ be the Bernoulli random variable, $Ber(\pi_n)$, which denotes the jth simulation result for the given sample size, n ($j = 1, \ldots, m$, $X_n = \sum_{j=1}^{m} Y_{n(j)}$). Thus, equation (3) is also denoted by $\pi_n = Pr(Y_{n(j)} = 1|n) = \Phi(\beta_0 + \beta_1 \sqrt{n})$.

Let Γ be a set of sample sizes in the simulations. For example, Γ = \{10, 15, 20, 25, 30, 35, 40, 45\} for $n_p = 28$. Thus, while n_p does not have to be a member of Γ, it must fall within the range of Γ. In this study, we apply model (3) to the simulation results, $\{\sqrt{n}, y_{n(j)}\}$ ($n \in \Gamma, j = 1, \ldots, m$). Let $\hat{\boldsymbol{\beta}}$ denote the MLE of the probit model parameter vector, $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$. The log likelihood, $\ell_m(\boldsymbol{\beta})$, and the expected information matrix, $I(\boldsymbol{\beta})$, are given by

$$\ell_m(\boldsymbol{\beta}) = \sum_{n \in \Gamma} \left[\log {}_m\mathrm{C}_{X_n} + X_n \log\{\Phi(\beta_0 + \beta_1 \sqrt{n})\} + (m - X_n) \log\{1 - \Phi(\beta_0 + \beta_1 \sqrt{n})\} \right] \tag{4}$$

and

$$I(\boldsymbol{\beta}) = m \sum_{n \in \Gamma} \frac{\{\phi(\beta_0 + \beta_1\sqrt{n})\}^2}{\Phi(\beta_0 + \beta_1\sqrt{n})\{1 - \Phi(\beta_0 + \beta_1\sqrt{n})\}} \begin{pmatrix} 1 & \sqrt{n} \\ \sqrt{n} & n \end{pmatrix},$$

respectively, where ${}_m\mathrm{C}_{X_n}$ denotes the number of X_n combinations from m elements, and $\phi(\cdot)$ is the probability density function of the standard normal distribution (see, e.g., Demidenko (2001)). Thus, the estimator of n_p based on the probit model is given by

$$n_p(\hat{\boldsymbol{\beta}}) = \left\lceil \nu_p(\hat{\boldsymbol{\beta}}) \right\rceil,$$

where $\nu_p(\hat{\boldsymbol{\beta}}) = (z_p - \hat{\beta}_0)^2/\hat{\beta}_1^2$. Under the assumption that $\nu_p(\hat{\boldsymbol{\beta}})$ follows a normal distribution, the approximate variance estimator of $\nu_p(\hat{\boldsymbol{\beta}})$ is given by $Var\{\nu_p(\hat{\boldsymbol{\beta}})\} \simeq \Delta^{\mathrm{T}} I(\hat{\boldsymbol{\beta}})^{-1}\Delta$, where

$$\Delta = \frac{\partial}{\partial \boldsymbol{\beta}} \nu_p(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \begin{pmatrix} -2(z_p - \hat{\beta}_0)/\hat{\beta}_1^2 \\ -2(z_p - \hat{\beta}_0)^2/\hat{\beta}_1^3 \end{pmatrix}.$$

Then, for a finite m, the approximate variance estimator of $n_p(\hat{\boldsymbol{\beta}})$ is given by

$$Var\{n_p(\hat{\boldsymbol{\beta}})\} = \sum_{n=1}^{\infty} n^2 Pr\left\{n_p(\hat{\boldsymbol{\beta}}) = n\right\} - \left[\sum_{n=1}^{\infty} n Pr\left\{n_p(\hat{\boldsymbol{\beta}}) = n\right\}\right]^2,$$

where

$$Pr\left\{n_p(\hat{\boldsymbol{\beta}}) = n\right\} = Pr\left\{n - 1 < \nu_p(\hat{\boldsymbol{\beta}}) \leq n\right\} \simeq \Phi\left[\frac{n - \nu_p(\boldsymbol{\beta})}{\sqrt{Var\{\nu_p(\hat{\boldsymbol{\beta}})\}}}\right] - \Phi\left[\frac{n - 1 - \nu_p(\boldsymbol{\beta})}{\sqrt{Var\{\nu_p(\hat{\boldsymbol{\beta}})\}}}\right].$$

In practice, it is sufficient to calculate the sum for the range of $n_p(\boldsymbol{\beta}) \pm n_p(\boldsymbol{\beta})/5$. Furthermore, the ceiling function would not have a significant effect on the results; that is $Var\{n_p(\hat{\boldsymbol{\beta}})\} \simeq Var\{\nu_p(\hat{\boldsymbol{\beta}})\}$, unless n_p is small.

If equation (3) holds, the estimated sample size has consistency unless $\pi_{n_p} = p$, similarly to the ordinary procedure. In addition, the finite sample bias for the estimated sample size is smaller than 0.5 under the assumption that equation (3) holds and $\nu(\hat{\boldsymbol{\beta}})$ follows a normal distribution. When the value of $\nu(\boldsymbol{\beta})$ is close to an integer (e.g., $\nu(\boldsymbol{\beta}) = 49.9$ or 49.1), the integer part of $\nu(\boldsymbol{\beta})$ is variable and, therefore, the convergence speed decreases considerably. The same is true for the ordinary procedure.

3 Effect of a model misspecification

Obviously, there exist situations where model (3) is not the true relationship between the sample size and the success probability. In this section, we investigate the effect of a model misspecification. Now, we assume that π_n is controlled by a parametric model with a parameter vector, $\boldsymbol{\gamma}$: $X_n \sim Bin(m, \pi_n(\boldsymbol{\gamma}))$. Under this assumption, we set the misspecified model, $Bin(m, \Phi(\beta_0 + \beta_1\sqrt{n}))$, which is described in Section 2. If we assume that for given $\boldsymbol{\gamma}$, $\hat{\boldsymbol{\beta}}$ converges in probability as $m \to \infty$ to a limit $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = (\beta_{\boldsymbol{\gamma}0}, \beta_{\boldsymbol{\gamma}1})^{\mathrm{T}}$, then $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is obtained by solving the equation $E_{\boldsymbol{\gamma}}\{\partial \ell_m(\boldsymbol{\beta})/\partial \boldsymbol{\beta}\} = \mathbf{0}$, where $E_{\boldsymbol{\gamma}}$ denotes the expectation under the true model (Cox, 1961). More specifically, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is obtained by solving the simultaneous equations

$$\sum_{n\in\Gamma} \phi(\beta_0 + \beta_1\sqrt{n}) \left\{ \frac{\pi_n(\boldsymbol{\gamma})}{\Phi(\beta_0 + \beta_1\sqrt{n})} + \frac{1 - \pi_n(\boldsymbol{\gamma})}{1 - \Phi(\beta_0 + \beta_1\sqrt{n})} \right\} = 0,$$

$$\sum_{n\in\Gamma} \sqrt{n}\phi(\beta_0 + \beta_1\sqrt{n}) \left\{ \frac{\pi_n(\boldsymbol{\gamma})}{\Phi(\beta_0 + \beta_1\sqrt{n})} + \frac{1 - \pi_n(\boldsymbol{\gamma})}{1 - \Phi(\beta_0 + \beta_1\sqrt{n})} \right\} = 0,$$

for β_0 and β_1. Since these equations cannot be solved explicitly, we solve them using the `nleqslv` function in `R` software. For given $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, the success probability based on the misspecified model is given by $\pi_n^*(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \Phi(\beta_{\boldsymbol{\gamma}0} + \beta_{\boldsymbol{\gamma}1}\sqrt{n})$. Two examples for $\pi_n(\boldsymbol{\gamma})$ follow. We do not consider the ceiling function in this section.

3.1 Case 1: Futility stopping design

The first example is a two-arms, parallel group, superiority, randomized clinical trial that contains an interim analysis for futility stopping. The distributions of groups 1 and 2 are $N(\mu_1, 1)$ and $N(\mu_2, 1)$, respectively. For simplicity, the variance for each group is assumed to be known. Let n denote the sample size per group for the final analysis. The aim of the trial is to demonstrate that $\eta = \mu_2 - \mu_1 > 0$. The final analysis is a two-sample Z-test with the both-tailed hypothesis and the significance level α. The interim analysis is conducted when $2rn$ subjects are completed ($0 < r < 1$), where the conditional power is calculated under the assumption that the estimated mean difference, $\hat{\eta}$, is the true value. The trial is terminated for futility if the calculated conditional power is less than ω; otherwise, the trial is continued. Thus, the overall power that accounts for the futility stopping, $\pi_n(\boldsymbol{\gamma})$, is

given by

$$\pi_n(\boldsymbol{\gamma}) = \Phi_2\left(\sqrt{\frac{rn}{2}}\eta - \frac{z_{1-\alpha/2} + \sqrt{1-r}z_\omega}{\sqrt{r} + \sqrt{1-r}}, \sqrt{\frac{n}{2}} - z_{1-\alpha/2}; \sqrt{r}\right),$$

where $\Phi_2(\cdot, \cdot; \rho)$ is the cumulative distribution function (CDF) of the standard bivariate normal distribution with correlation ρ. We now set $\alpha = 0.05$, $\gamma = 0.5$, and $r = 0.5$, and then η is set such that $\pi_n(\boldsymbol{\gamma}) = 0.8$ for $n = 100$. We also set $\Gamma = \{k | k = 5 + 5l \leq 200, l = 0, 1, 2, \ldots\}$. Under these conditions, $\boldsymbol{\beta}_\gamma$ and $\pi_n^*(\boldsymbol{\beta}_\gamma)$ are calculated.

Figure 1(a) shows the relationship between n and $\pi_n(\boldsymbol{\gamma})$ (solid line) or $\pi_n^*(\boldsymbol{\beta}_\gamma)$ (dashed line). The overall power curve based on the misspecified model seems to be sufficiently close to the true power curve.

3.2 Case 2: Equivalence trial

The second example is a two-arms, parallel group, randomized clinical trial that aims to demonstrate the equivalence of two treatments. The distribution of the endpoint for groups 1 and 2 are $N(\mu_1, 1)$ and $N(\mu_2, 1)$, respectively. For simplicity, the variance for each group is also assumed to be known here. Let n, η, and $C > 0$ denote the sample size per

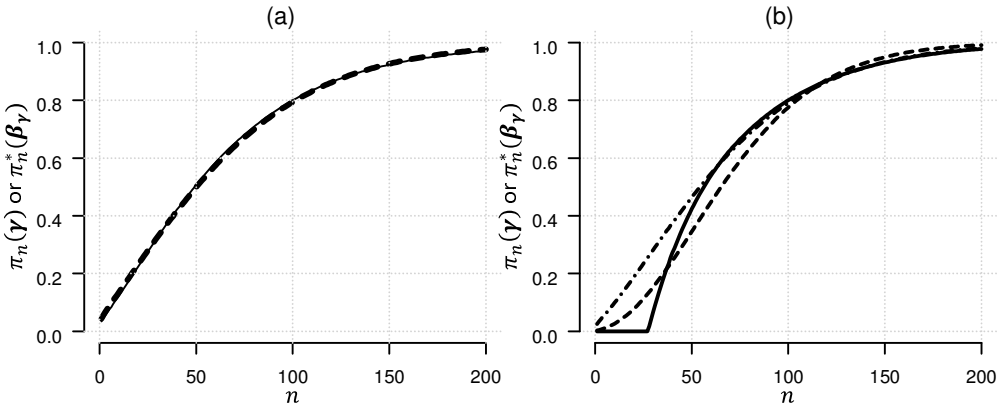

Figure 1: Comparison of the true and probit model-based success probability curves. (a) Results for case 1. Solid line: true power curve; dashed line: probit model-based power curve. (b) Results for case 2. Solid line: true power curve; dashed line: probit model-based power curve for $\Gamma = \{k | k = 5 + 5l \leq 200, l = 0, 1, 2, \ldots\}$; dotted-dashed line: probit model-based power curve for $\Gamma = \{k | k = 50 + 5l \leq 200, l = 0, 1, 2, \ldots\}$.

group, mean difference $(\mu_2 - \mu_1)$, and equivalence margin, respectively. If the $100(1-\alpha)\%$ confidence interval for η is contained within the interval $[-C, C]$, equivalence is declared. Thus, the probability that equivalence is established is given by

$$\pi_n(\boldsymbol{\gamma}) = \max\left[0, \Phi\left\{\sqrt{n/2}(C-\eta) - z_{1-\alpha/2}\right\} + \Phi\left\{\sqrt{n/2}(C+\eta) - z_{1-\alpha/2}\right\} - 1\right].$$

We now set $\alpha = 0.05$, $\eta = C/4$, and C is set such that $\pi_n(\boldsymbol{\gamma}) = 0.8$ for $n = 100$. We also set $\Gamma = \{k | k = 5 + 5l \leq 200, l = 0, 1, 2, \ldots\}$. Under these conditions, $\boldsymbol{\beta}_\gamma$ and $\pi_n^*(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ are calculated.

Figure 1(b) shows the relationship between n and $\pi_n(\boldsymbol{\gamma})$ (solid line) or $\pi_n^*(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ (dashed line). Here, the probability curve based on the misspecified model does not fit the true curve well.

Since our motivation is to estimate n_p, it is sufficient to approximate the true curve near the point of (n_p, p). Therefore, we narrow the range of Γ: $\Gamma = \{k | k = 50 + 5l \leq 200, l = 0, 1, 2, \ldots\}$, then recalculate $\pi_n^*(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ and plot it as the dotted-dashed line in Figure 1(b). The curve for the narrower Γ seems to fit the true curve sufficiently in the range of $50 \leq n \leq 200$.

3.3 Range of Γ

From the previous results, it is suggested that the estimated curve based on the probit model might not fit the true model well if the range of Γ is too wide. However, the efficiency of the estimation would decrease if the range of Γ is too narrow. Therefore, we investigate the optimal range of Γ for general situations on the basis of cases 1 and 2. We denote $\Gamma = \{k | k = n_1 + 5l \leq n_2, l = 0, 1, 2, \ldots\}$. The parameter settings other than Γ and p are the same as the settings described in Sections 3.1 and 3.2. Then, n_1 and n_2 are set such that π_{n_1} and π_{n_2} become the specified values. We set $p = 0.8$ or 0.9, $n_p = 100$, and $\pi_{n_1} = 0.2$, 0.4, or 0.6. We also set $\pi_{n_2} \in [0.8, 1]$ for $p = 0.8$, and $\pi_{n_2} \in [0.9, 1]$ for $p = 0.9$. For each parameter setting, the bias of the estimator for n_p based on the probit model is evaluated in percentage terms: $100\{n_p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) - n_p\}/n_p$.

Figure 2 shows the bias evaluation results. The bias decreases as π_{n_1} increases up to p, and seems acceptable for all cases when $\pi_{n_1} = 0.6$. For π_{n_2}, we evaluate the results

for $\pi_{n_1} = 0.6$ only. The bias reaches a minimum around $\pi_{n_2} = 0.9$ and 0.95 for $p = 0.8$ and 0.9, respectively. In practice, there would be few situations such that the success probability curve is not smooth in the range of [0.6,0.95], other than for some exact test methods. These results indicate that the settings of $(\pi_{n_1}, \pi_{n_2}) = (0.6, 0.9)$ for $p = 0.8$, and $(\pi_{n_1}, \pi_{n_2}) = (0.6, 0.95)$ for $p = 0.9$ are reasonable. Although we evaluate the bias for fixed $n_p = 100$, the above discussion for the percentage bias does not change if the value of n_p changes, because only the scale of the horizontal axis of Figure 1 changes.

3.4 Proposed simulation procedure

On the basis of the above discussion, the proposed simulation procedure based on the probit model is given as follows:

1. Set the number of simulations, m, and start the simulations from $n = 10$.

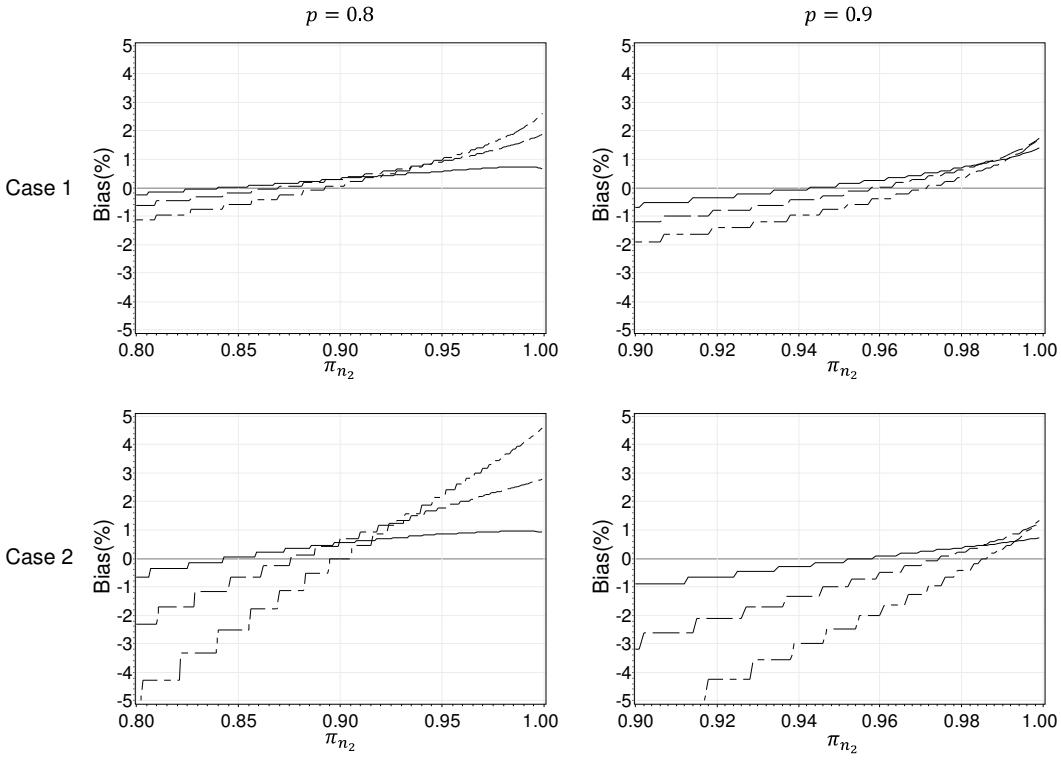

Figure 2: Bias evaluation results based on the probit model for cases 1 and 2, $p = 0.8, 0.9$. dotted-dashed line: $\pi_{n_1} = 0.2$; dashed line: $\pi_{n_1} = 0.4$; solid line: $\pi_{n_1} = 0.6$.

2. Determine the step size, s, for n as follows:

   (a) If $\hat{\pi}_{10(m)} > 0.5$, then restart the simulations from $n = 2$ and set $s = 1$.

   (b) If $0.3 < \hat{\pi}_{10(m)} \leq 0.5$, then set $s = 2$.

   (c) If $\hat{\pi}_{10(m)} \leq 0.3$, then set $s = 5$.

3. Conduct simulations, increasing n by s until $\hat{\pi}_{n(m)}$ exceeds 0.9 (0.95) twice for $p = 0.8$ (0.9).

4. Set $\Gamma = \{k | k = n_1 + sl \leq n_2, l = 0, 1, 2, \ldots\}$. Here, n_1 is the smallest sample size such that $\hat{\pi}_{n(m)} > 0.6$ and n_2 is the sample size at which the simulations terminate.

5. Apply the probit model in formula (3) to the simulation data $\{n, y_{n(j)}\}, n \in \Gamma, j = 1, \ldots, m$, and then obtain the MLE, $\hat{\boldsymbol{\beta}}$.

6. Obtain the estimated sample size as $\lceil (z_p - \hat{\beta}_0)^2 / \hat{\beta}_1^2 \rceil$.

Here, $\hat{\pi}_{n(m)}$ is the simple MLE given in section 2.1. Step 2 (a) and (b) prevent the number of members of Γ from becoming too small. One addable option is increasing the step size. For example, if $\hat{\pi}_{100} < 0.3$, then n_p would be larger than 300 and, therefore, it would be better to increase s to 10 or 20.

Note that the proposed procedure can be applied with other statistical models such as the logistic model or a nonparametric regression. However, in this study, we use the probit model to maintain theoretical consistency. The determination of m is discussed in Sections 4 and 5.

4 Comparison of ordinary and proposed simulation procedures

4.1 Theoretical comparison

In this section, we compare the performances of the ordinary and proposed simulation procedures using the theoretical approach summarized in Section 2. We set the situation

Table 1: Minimum number of simulations such that CV$<$ 1% for the ordinary and proposed simulation procedures.

Procedure	$n_p = 50$	$n_p = 200$	$n_p = 1000$
Ordinary	12596	5175	2829
Proposed	1884	382	79

such that a two-sample t-test is conducted for two normal populations with known variance, 1. Then, π_n is given by $\pi_n = \Phi(z_{1-\alpha/2} + (\eta/\sqrt{2})\sqrt{n})$, where η is the mean difference between the two populations. We set the target power and the sample size per group, $p = 0.8$ and $n_p = 50, 200, 1000$, respectively, and then, η is set such that $\pi_{n_p-0.5} = p = 0.8$; $\eta = \sqrt{2/(n_p - 0.5)}(z_{1-\alpha/2}+z_p)$. We use the condition that $\pi_{n_p-0.5} = p$ so that the estimator based on the ordinary procedure has consistency. The range of m is from 100 to 10000. The proposed procedure is defined for $\Gamma = \{k|n = n_1 + 5l \leq n_2, l = 0, 1, 2, \ldots\}$, where n_1 and n_2 are the smallest sample sizes such that $\pi_n > 0.6$ and $\pi_n > 0.9$, respectively. The ordinary procedure is given in Section 2.1. We calculate the bias for the ordinary procedure as $100\{E(\hat{n}_{p(m)}) - n_p\}/n_p$ and the coefficient of variation (CV) for the ordinary and proposed procedures as $100\sqrt{Var\{\hat{n}_{p(m)}\}}/n_p$ and $100\sqrt{Var\{n_p(\hat{\boldsymbol{\beta}})\}}/n_p$, respectively.

Figure 3 shows the bias for the ordinary simulation procedure. For $n_p > 200$, the size of the bias is larger than 3%, even when $m = 1000$. Figure 4 shows the CVs for the ordinary and proposed procedures. Furthermore, the minimum numbers of simulations (m) such that CV$<$ 1% for the ordinary and proposed procedures are given in Table 1. The precision of the proposed procedure is much higher than that for the ordinary procedure. If we conduct simulations based on the proposed procedure for $m = 1000$, the variation of the simulation results range by roughly $\pm$2% with a 95% confidence level. If we want to estimate the sample sizes based on the ordinary procedure with the same precision as the proposed procedure, we would have to set m to be more than five times that of the proposed procedure. The discrepancy in the precision between the two procedures becomes larger as n_p increases.

4.2 Simulation study

In this section, we discuss the designs and results of the two simulation studies we have conducted.

Simulation 1. Simulation 1 aims to confirm the theoretical comparison results described in Section 4.1 for finite m. We set $n_p = 50$ and $m = 100$, 300, 1000, 3000, and 10000.

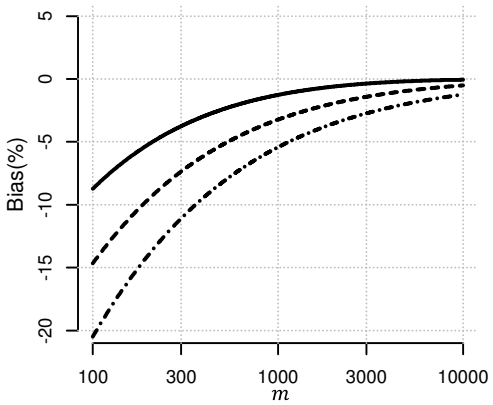

Figure 3: Relationship between m and the bias in the estimated sample size for the ordinary simulation procedure. Solid line: $n_p = 50$; dashed line: $n_p = 200$; dotted-dashed line: $n_p = 1000$. m is shown using the common logarithm scale.

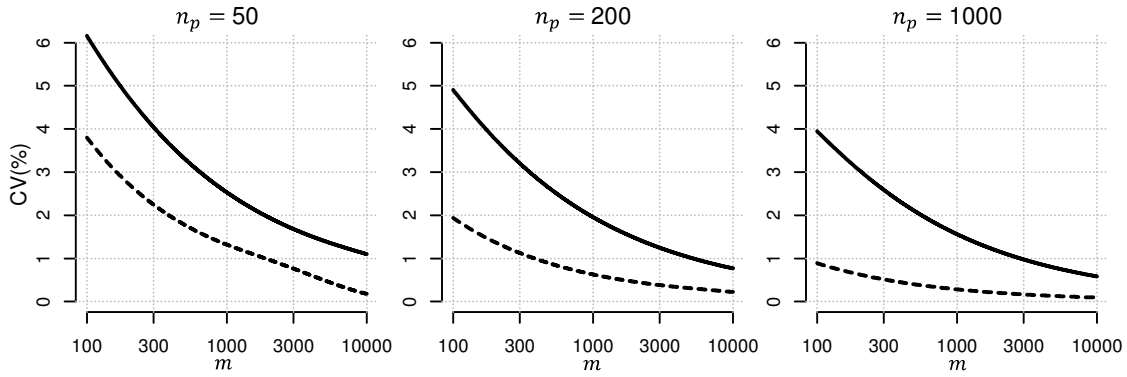

Figure 4: Relationship between m and the coefficient of variation of the estimated sample size. Solid line: ordinary simulation procedure; dashed line: proposed simulation procedure. m is shown using the common logarithm scale.

The starting sample sizes for the ordinary and proposed procedures were both set to 10. The other settings were the same as those in Section 4.1. For each value of m, the sample sizes were estimated based on the ordinary and proposed procedures. We repeated these processes 1000 times and calculated the mean and the standard deviation of the estimated sample sizes as the simulated expectation and standard error (SE), respectively, for each m and each procedure.

Figure 5 shows the simulation and the corresponding theoretical results. Although there was a slight discrepancy between the simulation and the theoretical results for the case where $m = 100$, the simulation results were sufficiently close to the theoretical results for almost all the situations. On the other hand, the precision for the ordinary procedure was much lower than that of the proposed procedure, and the bias for small m was large, which were the results predicted by the theoretical comparison.

Simulation 2. Simulation 2 compares the two simulation procedures in the complicated setting where no sample size formula is available and the computational cost is high. The

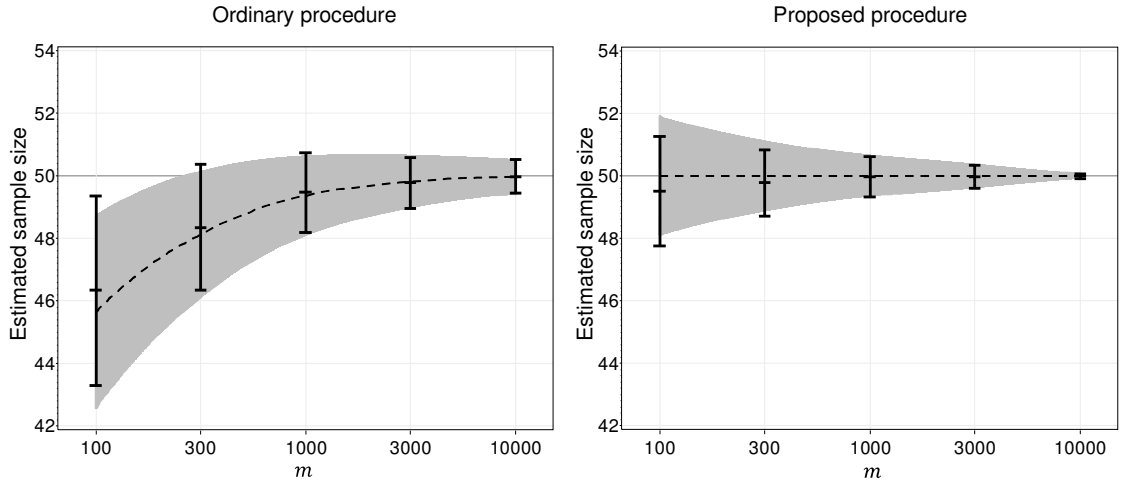

Figure 5: Results of simulation 1. Error bar plot: mean ± standard deviation for simulated sample size estimates; dotted line: theoretical expectation for estimated sample size; gray area: range of theoretical mean ± theoretical SE. Further, m is shown using the common logarithm scale.

Table 2: Estimated sample size for Simulation 2. Cor: correlation structure, q: missing probability at last occasion, ρ: correlation parameter for specified correlation structure, d: efficacy parameter.

Settings				Ordinary procedure			Proposed procedure		
				m			m		
Cor	q	ρ	d	300	1000	3000	300	1000	3000
AR(1)	0.2	0.4	0.3	284	294	301	296	302	301
AR(1)	0.2	0.4	0.4	161	167	174	168	170	170
AR(1)	0.2	0.4	0.5	106	110	112	110	111	109
AR(1)	0.2	0.8	0.3	214	223	223	228	224	223
AR(1)	0.2	0.8	0.4	122	125	127	126	129	127
AR(1)	0.2	0.8	0.5	84	78	83	84	82	82
AR(1)	0.4	0.4	0.3	506	532	527	540	542	542
AR(1)	0.4	0.4	0.4	291	299	305	305	306	306
AR(1)	0.4	0.4	0.5	184	197	196	200	197	199
AR(1)	0.4	0.8	0.3	347	350	361	365	366	364
AR(1)	0.4	0.8	0.4	202	206	207	212	208	209
AR(1)	0.4	0.8	0.5	132	137	136	136	136	136
CS	0.2	0.4	0.3	229	245	247	247	244	245
CS	0.2	0.4	0.4	134	140	138	139	141	141
CS	0.2	0.4	0.5	84	88	94	90	92	90
CS	0.2	0.8	0.3	96	101	101	99	103	103
CS	0.2	0.8	0.4	61	60	60	58	60	59
CS	0.2	0.8	0.5	41	40	41	39	40	40
CS	0.4	0.4	0.3	390	422	424	428	433	433
CS	0.4	0.4	0.4	222	241	247	244	246	246
CS	0.4	0.4	0.5	160	156	157	162	161	161
CS	0.4	0.8	0.3	158	171	172	173	173	174
CS	0.4	0.8	0.4	96	101	104	102	101	101
CS	0.4	0.8	0.5	66	66	69	67	67	68

Table 3: Computational time (hour) for Simulation 2.

Procedure	$m = 300$	$m = 1000$	$m = 3000$
Ordinary	5.1	19.6	64.3
Proposed	2.9	10.7	33.2

planned trial design was randomized, two-group, placebo-controlled, parallel design. The efficacy endpoints were observed at baseline ($j = 0$) and the post-treatment four occasions

($j = 1, 2, 3, 4$), where the primary occasion was the last one. The means for the placebo and test drug groups were assumed as $\{0, 0.05, 0.10, 0.15, 0.20\}$ and $\{0, (0.2-d)/4, 2(0.2-d)/4, 3(0.2-d)/4, 0.2-d\}$, respectively ($d = 0.3$: minimum requirement, 0.4: moderate, 0.5: desirable). The standard deviation for all groups and occasions were set as 1. We set the correlation structure for occasions as the compound symmetry (CS) and first-order autoregression (AR(1)) structures, where the values of the correlation parameter, ρ, were set as 0.4 or 0.8. The missing structure was set as monotone and missing at random. The missing probability was modeled by $logit\{Pr(R_j = 1)\} = Int + x_{j-1}$, where x_j was the value of the endpoint at the jth occasion and R_j was the indicator random variable, such that $R_j = 1$ when x_j was missing, otherwise, $R_j = 0$ ($j = 1, 2, 3, 4$). We also set $R_0 = 0$ (i.e., no missing values at the baseline). Then, Int was calculated such that the missing proportion of the combined group at the last occasion became $100q\%$ ($q = 0.2$, 0.4). The outcome at the last occasion was analyzed using an analysis of covariance, including the treatment group as a factor and the baseline observation as a covariate. The missing values were imputed by the control-based pattern imputation procedure (Ratitch

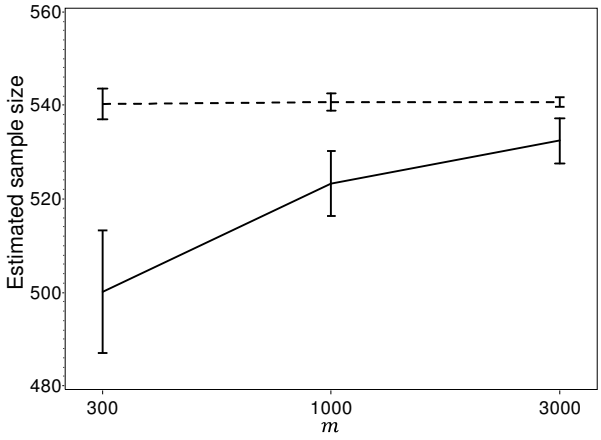

Figure 6: Results of simulation 2 (correlation structure: AR(1), missing probability (q): 0.4, correlation parameter (ρ): 0.4, efficacy parameter (d): 0.3). Error bar plot: mean $\pm$ standard deviation for simulated sample size estimates. Solid line: ordinary simulation procedure; dashed line: proposed simulation procedure. m is shown at the common logarithm scale.

et al., 2013) with an imputation number of 10. We calculated the sample sizes for the 24 settings of the primary analysis (correlation structure: {CS, AR(1)} , missing probability: $q = 0.2, 0.4$, correlation parameter: $\rho = 0.4, 0.8$, efficacy: $d = 0.3, 0.4, 0.5$) using the ordinary and proposed simulation procedures for $m = 300$, 1000, and 3000. The step size for the ordinary procedure was set as 10 when $\hat{\pi}_{n(m)} \leq 0.3$, 5 when $0.3 < \hat{\pi}_{n(m)} \leq 0.5$, and 1 when $\hat{\pi}_{n(m)} \geq 0.5$. We set the target power as $p = 0.8$. We measured the simulation run-time of the primary analysis for each value of m and each procedure (computation environment: CPU: Intel(R) Xeon(R) CPU E5-2637 v3 @ 3.50 GHz ; memory: 32.0GB, Software: SAS(R) 9.4; MI procedure). These simulations were conducted in parallel with 10 CPU cores and the computational time was calculated as the summation of the 10 measurements.

Furthermore, we conducted 100 simulations (e.g., 3000×100 times for $m = 3000$) for the above setting, which provided the maximum sample size. Then, we calculated the means and standard deviations of the 100 estimated sample sizes for each procedure and each value of m in order to evaluate the accuracy and precision of the two simulation procedures.

Table 2 shows the estimated sample size for each value of m and simulation procedure. The proposed procedure gave almost same estimated values for all values of m, while the sample sizes from the ordinary procedure for $m = 300$ were unstable and would be seriously biased, especially when the estimated value was large. Table 3 shows the computation time for Simulation 2. The computation time for the ordinary method was about twice as long as that of the proposed procedure for the same values of m. Figure 6 shows the simulation results for the 100 estimated sample sizes for the setting that gave the maximum sample size (correlation structure: AR(1), $q = 0.4$, $\rho = 0.4$, $d = 0.3$). The mean sample sizes for the ordinary procedure were considered to be underestimated, even for $m = 3000$. The precision of the proposed procedure for $m = 100$ was higher than that of the ordinary procedure for $m = 3000$. This suggests that the accuracy and precision of the ordinary procedure were much lower than those of the proposed procedure.

5 Discussion

In this paper, we summarized the theoretical aspect of the sample size calculation based on statistical simulations, and proposed a simulation procedure based on the probit model. It was found that the proposed procedure allowed us to estimate the required sample size with higher accuracy and precision than the ordinary simulation procedure even when the probit model was not the true structure of the power curve. Our simulation also showed that the computational time for sample size calculation with our proposed procedure was much shorter than that for the ordinary procedure. Especially when n_p was moderate to large, the time for the proposed procedure would become over 10 times shorter than that for the ordinary procedure because the precision of the ordinary procedure for $m = 3000$ is lower than that of the proposed procedure for $m = 300$ from the results for $n_p = 200$ in Figure 4.

Now, we discuss the setting of the simulation size, m. If we set $m = 1000$, the SE for estimated sample size would become lower than $\max(1, 0.01n_p)$ for the proposed procedure, and such precision would be acceptable when we evaluate and compare many scenarios for planning a clinical trial. When the objective is that multiple scenarios are roughly compared, it might be sufficient to set $m = 200$–500 because sample size would be estimated with little bias and about 5% precision. On the other hand, the sample size estimator based on the ordinary simulation procedure might be biased even if $m \geq 3000$ especially for large n_p. When we calculate the sample size for the main scenario in the protocol description, it would be desirable to set an adequately large value of m (e.g., $m = 10000$).

Although we applied only one ordinary procedure, some other "ordinary" procedures might be available. For example, the step size can be changed, the simulation can be started from large sample size and stepped down, and the bisection method would be available. Nevertheless, our proposed approach would still have the substantial advantages of high precision and accuracy due to the statistical model fitting over other ordinary algorithm-based approaches because the statistical model based procedure uses the simulation data efficiently. These advantages would also result in the reduction of the computation time.

As our manuscript provides only the framework of the simulation procedure based on some statistical models, the proposed procedure may still have room for improvement. For

example, Γ and m can be set more adaptively in response to the intermediate simulation results. Also, our proposed procedure assumes equal sample sizes for multiple groups. We can introduce an allocation ratio parameter for considering unequal sample size in our procedure. This proposed procedure can also be applied to calculate the effect size such that the success probability for a fixed sample size attains a fixed level. While there are less opportunities where the effect size is calculated, one might want to calculate the detectable effect size for a fixed sample size. Although we only focused on statistical hypothesis test with significance level of 0.05 in sections 3 and 4, this procedure would be useful for situations where the success probability function for amount of information, n, shows smooth and monotone increasing curve at least for the range of $p \in [0.6, 1]$. Such situations would be very common. Although we only focused on the probit model, simulation procedures based on some other statistical models such as semiparametric models would be the subject of future investigation.

Nevertheless, as our proposed simulation procedure applying statistical models to the simulation data would help to reduce the computational time required to conduct clinical trial simulations dramatically, it contributes to the cost reduction of the clinical trial designing and/or the ease of wide-range parameter settings for the clinical trial simulations.

References

Bretz, F., H. Schmidli, F. Koenig, A. Racine, and W. Maurer (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. *Biometrical Journal 48*, 623–634.

Burton, A., D. G. Altman, P. Royston, and R. L. Holder (2006). The design of simulation studies in medical statistics. *Statistics in Medicine 25*, 4279–4292.

Chen, Y. H. J., D. L. DeMets, and K. K. G. Lan (2004). Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine 23*, 1023–1038.

Chow, S. C. and M. Chang (2011). *Adaptive Design Methods in Clinical Trials, 2nd edition.* Boca Raton: Chapman & Hall/CRC.

Cox, D. R. (1961). Tests of separate families of hypotheses. In *Proceedings of the 4th Berkely Symposium 1*, pp. 105–123. University of California Press.

Cui, L., H. M. J. Hung, and S. J. Wang (1999). Modification of sample size in group sequential clinical trials. *Biometrics 55*, 853–857.

Demidenko, E. (2001). Computational aspects of probit model. *Mathematical Communications 6*, 233–247.

Julious, S. A. (2004). Tutorial in biostatistics: sample sizes for clinical trials with normal data. *Statistics in Medicine 23*, 1921–1986.

Julious, S. A. and M. J. Campbell (2012). Tutorial in biostatistics: sample sizes for parallel group clinical trials with binary data. *Statistics in Medicine 31*, 2904–2936.

Maca, J., S. Bhattacharya, V. Dragalin, P. Gallo, and M. Krams (2006). Adaptive seamless phase II/III designs-background, operational aspects, and examples. *Drug Information Journal 40*, 463–473.

Ratitch, B., M. O'Kelly, and R. Tosiello (2013). Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. *Pharmaceutical Statistics 12*, 337–347.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons, Inc.

Schoenfeld, D. A. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics 39*, 499–503.

Thall, P. F., J. K. Wathen, B. N. B. R. E. Champlin, L. H. Baker, and R. S. Benjamin (2003). Hierarchical bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in Medicine 22*, 763–780.

Tighiouart, M., A. Rogatko, and J. S. Babb (2005). Flexible Bayesian methods for cancer phase I clinical trials: dose escalation with overdose control. *Statistics in Medicine 24*, 2183–2196.

Zhang, J. and E. Pulkstenis (2016). Sample size and power of survival trials in group sequential design with delayed treatment effect. *Statistics in Biopharmaceutical Research 8*, 268–275.

Zhu, H. (2017). Sample size calculation for comparing two Poisson or megative binomial rates in noninferiority or equivalence trials. *Statistics in Biopharmaceutical Research 9*, 107–115.