Published in final edited form as:

Clin Trials. 2010 August ; 7(4): 343–353. doi:10.1177/1740774510373629.

Calculating Sample Size in Trials Using Historical Controls

Song Zhang

Department of Clinical Sciences UT Southwestern Medical Center Dallas, TX

Jing Cao

Department of Statistical Science Southern Methodist University Dallas, TX

Chul Ahn

Department of Clinical Sciences UT Southwestern Medical Center Dallas, TX

Abstract

Background—Makuch and Simon [1] developed a sample size formula for historical control trials. When assessing power, they assumed the true control treatment effect to be equal to the observed effect from the historical control group. Many researchers have pointed out that the M-S approach does not preserve the nominal power and type I error when considering the uncertainty in the true historical control treatment effect.

Purpose—To develop a sample size formula that properly accounts for the underlying randomness in the observations from the historical control group.

Methods—We reveal the extremely skewed nature in the distributions of power and type I error, obtained over all the random realizations of the historical control data. The skewness motivates us to derive a sample size formula that controls the percentiles, instead of the means, of the power and type I error.

Results—A closed-form sample size formula is developed to control arbitrary percentiles of power and type I error for historical control trials. A simulation study further demonstrates that this approach preserves the operational characteristics in a more realistic scenario where the population variances are unknown and replaced by sample variances.

Limitations—The closed-form sample size formula is derived for continuous outcomes. The formula is more complicated for binary or survival time outcomes.

Conclusions—We have derived a closed-form sample size formula that controls the percentiles instead of means of power and type I error in historical control trials, which have extremely skewed distributions over all the possible realizations of historical control data.

Keywords

clinical trial design; sample size; historical controls; percentiles of type I error and power

1 Introduction

Randomized clinical trials (RCT) have become the gold standard in comparing the effects between treatments. Despite the rigorous scientific basis, there are situations where RCTs are infeasible due to concerns of ethics, patient preference, cost, and regulatory acceptability. For example, the resources required by an RCT might be prohibitive for some

phase II trials which are intended to obtain preliminary data on the effectiveness of a new treatment [2]. Another example is that when evidence already exists showing the superiority of a new treatment over the standard one, it might be unethical for a RCT to assign patients to a potentially inferior treatment. One solution is to use a historical control trial (HCT), where the experimental therapy is compared with a control therapy (referred to as historical control or HC) that has been evaluated in a previously conducted trial. Because an HCT can be smaller in size and easier to conduct, it has been widely applied in clinical research [3, 4, 5, 6, 7, 8, 9].

Makuch and Simon [1] developed a sample size formula for HCTs with a binary outcome. In power calculation, they assumed that the observed response rate from the HC group was the true control response rate. Their formula was based on the two-sample test statistic employed in RCTs but the power calculation only accounted for the sampling variability in the experimental group. The sample size solution was obtained through a numerical search. Using a similar idea, Dixon and Simon [10] provided a sample size formula for HCTs with exponential survival outcomes. Chang et al. [11] presented a two-stage design for phase II clinical trials with HC and continuous outcomes. More discussions about the HCT sample size calculation can be found in [12] and [13].

The estimated sample size for an HCT is usually much smaller than that required by an RCT. Lee and Tseng [14] pointed out that the sample size reduction in HCT is largely unjustified due to the strong assumption that the observed historical control response rate is equal to the true control response rate. They proposed a uniform power method to control the expected power, taking into account the uncertainty in the HC response rate. The resulting sample size is closer to the RCT sample size than the one based on Makuch and Simon's method (M-S). Korn and Freidlin [15] compared three approaches to HCT design: M-S approach, RCT approach, and one-sample approach (based on a one-sample test that the experimental treatment effect is greater than the observed HC treatment effect). The authors suggested to adopt the RCT approach because it preserves the unconditional power over the random HC observations.

In this study, we investigate the sample size calculation for HCT with continuous outcomes, accounting for uncertainty caused by the unknown true HC treatment effect. We provide a unified framework for the M-S, RCT and one-sample approaches, where they are shown to either control the mean or the median of the random power and type I error, obtained over all the possible realizations of the HC data given the true HC effect. We further demonstrate through simulation that the distributions of power and type I error are extremely skewed. This extreme skewness leads to undesirable properties of sample sizes calculated to control the means of power and type I error. One revealing example in our simulation is that with the mean power controlled at 0.8, a slight decrease in the mean type I error from 0.06 to 0.05 leads to a drastic increase in sample size from 286 to 487. This observation motivates us to develop a sample size formula that controls the percentiles, instead of the means, of the random power and type I error. To our knowledge, it is the first study in HCT design to demonstrate the extreme skewness in the distributions of power and type I error, and to estimate sample size based on the percentiles of power and type I error. It provides researchers a sensible way to assess the risk in an HCT. The proposed formula has a closed form, which can be easily computed using a scientific calculator.

The rest of this paper is organized as follows. In Section 2 we review the three different approaches (M-S, RCT, and one-sample) to sample size calculation in HCT under a unified framework. A simulation study is conducted to demonstrate the extreme skewness in the distributions of power and type I error. In Section 3 we present a sample size formula to control arbitrary percentiles of power and type I error. We evaluate its performance through

simulation in two scenarios: an ideal scenario (population variances known) and a more realistic scenario (population variances unknown). In Section 4 we provide a real application of the proposed method. The final section is devoted to discussion.

2 A Unified Framework

We briefly review the M-S, RCT, and one-sample approaches to HCT sample size calculation. Suppose in a clinical trial we compare the outcomes between an experimental group and an HC group. The outcome variable is continuous following a normal

distribution. Let $Y_1, \dots, Y_m \stackrel{iid}{\sim} N(\theta_0, \sigma_0^2)$ be the m observations from the HC group, and $X_1, \dots, X_n \stackrel{iid}{\sim} N(\theta_1, \sigma_1^2)$ be the n observations from the experimental group. We define $Y = \{Y_1, \dots, Y_m\}$ and $X = \{X_1, \dots, X_n\}$. The variances σ_0^2 and σ_1^2 are assumed to be known. With null hypothesis $H_0: \theta_1 = \theta_0$ and alternative hypothesis $H_1: \theta_1 > \theta_0$, the standard test statistic is

$$Z(X, Y) = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_1^2/n + \sigma_0^2/m}}, \quad (1)$$

where $\bar{X} = \sum_{i=1}^n X_i/n$ and $\bar{Y} = \sum_{j=1}^m Y_j/m$ are the sample means from the two groups. Given type I error α , power $1 - \beta$, m , σ_0^2 , σ_1^2 , and difference in treatment effects $\Delta = \theta_1 - \theta_0$, the sample size n is obtained by solving

$$P\left(\frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_1^2/n + \sigma_0^2/m}} > z_{1-\alpha}\right) = 1 - \beta, \quad (2)$$

where $z_{1-\alpha}$ is the $100(1 - \alpha)$ th percentile of the standard normal distribution. Here and in the rest of the paper we do not differentiate the estimated sample size or the solution to the sample size equations, with the understanding that the final sample size is the smallest integer greater than or equal to the solution.

In the M-S approach, the Y_j s from the HC group are considered not subject to sampling variability because they have been observed before the clinical trial. This consideration leads to the following manipulation of (2),

$$\begin{aligned} & P\left(\frac{(\bar{X} - \bar{Y}) - (\Delta + \theta_0 - \bar{Y})}{\sqrt{\sigma_1^2/n}} > \frac{z_{1-\alpha} \sqrt{\sigma_1^2/n + \sigma_0^2/m} - \frac{\Delta + \theta_0 - \bar{Y}}{\sigma_1^2/n}}{\sqrt{\sigma_1^2/n}}\right) \\ &= 1 - \Phi\left(\frac{z_{1-\alpha} \sqrt{\sigma_1^2/n + \sigma_0^2/m} - \frac{\Delta + \theta_0 - \bar{Y}}{\sigma_1^2/n}}{\sqrt{\sigma_1^2/n}}\right) = 1 - \beta, \end{aligned} \quad (3)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Thus we find n by solving

$$\frac{z_{1-\alpha} \sqrt{\sigma_1^2/n + \sigma_0^2/m}}{\sigma_1^2/n} - \frac{\Delta + \theta_0 - \bar{Y}}{\sqrt{\sigma_1^2/n}} = z_\beta. \quad (4)$$

Since the true HC effect (θ_0) is usually unknown, it is cancelled out in the equation by assuming $\theta_0 = \bar{Y}$. This is a strong assumption especially when the number of HC observations is limited. Traditionally Equation (4) has been solved through a numeric search [11]. Here we present a closed-form solution and a sufficient and necessary condition for its existence.

Theorem 1. Define $a = z_{1-\alpha}^2 \sigma_0^2 / (m \sigma_1^2) - \Delta^2 / \sigma_1^2$, $b = 2\Delta z_{1-\beta} / \sqrt{\sigma_1^2}$, and $c = z_{1-\alpha}^2 - z_{1-\beta}^2$. In clinical trials we usually specify α and β such that $\alpha \leq \beta$. Equation (4) has a unique sample size solution if and only if $\Delta > z_{1-\alpha} \sqrt{\sigma_0^2/m}$, and the solution is

$$n_0 = \left(\frac{-b - \sqrt{b^2 - 4ac}}{2a} \right)^2. \quad (5)$$

Proof. See Appendix A.1.

Theorem 1 helps researchers avoid time-consuming numerical search. It also points out a potential pitfall in the M-S approach where too small an assumed difference in the treatment effects ($\Delta \leq z_{1-\alpha} \sqrt{\sigma_0^2/m}$) would lead to no solution for the sample size.

In the one-sample approach, the hypotheses are specified as $H_0: \theta_1 = \bar{Y}$ and $H_1: \theta_1 > \bar{Y}$ based on the assumption that $\theta_0 = \bar{Y}$. The one-sample test statistic $Z_1(X, Y) = (\bar{X} - \bar{Y}) / \sqrt{\sigma_1^2/n}$ is employed. The sample size estimate is $n_1 = \sigma_1^2 (z_{1-\alpha} + z_{1-\beta})^2 / \Delta^2$.

In the RCT approach, the HC group is treated as a regular control arm in an RCT. The sample size is estimated by $n_2 = \sigma_1^2 (z_{1-\alpha} + z_{1-\beta})^2 m / [m \Delta^2 - (z_{1-\alpha} + z_{1-\beta})^2 \sigma_0^2]$, which is based on a two-sample test.

The three approaches produce drastically different sample size estimates. For example, given $m = 80$, $\sigma_0^2 = \sigma_1^2 = 1$, $\Delta = 0.3$, $\alpha = 0.05$, and $1 - \beta = 0.8$, the estimated sample sizes are $n_0 = 144$, $n_1 = 69$, and $n_2 = 487$, respectively. The formulas of n_0 , n_1 and n_2 do not depend on the specific values of HC sample mean ($\bar{Y}$) or true mean (θ_0). The test statistics, however, are calculated based on the HC sample mean. In a particular study, the unknown difference between $\bar{Y}$ and θ_0 has a great impact on the realized power and type I error. We conduct Simulation 1 to compare the performance of n_0 , n_1 and n_2 . Details of the simulation algorithm are presented in Appendix A.2.

The realization of a particular HC data ($Y^{(k)}$) leads to a conditional power ($q_v^{(k)}$) and a conditional type I error ($h_v^{(k)}$) under sample size n_v . Over the random realizations of Y , we have a random power, q_v , and a random type I error, h_v . The distributions of q_v and h_v

provide a global view of the variability in power and type I error for HCTs given an unknown HC treatment effect.

Without loss of generality, we set the true HC effect at $\theta_0 = 0$. We also assume $m = 80$, $\sigma_0^2 = \sigma_1^2 = 1$, $\Delta = 0.3$, $\alpha = 0.05$, and $1 - \beta = 0.8$. Figure 1 shows the results of Simulation 1. The graphs in the first column indicate that both the conditional power and type I error decrease monotonically as the difference between the observed and true HC treatment effect, $(\bar{Y}^{(k)} - \theta_0)$, increases. Table 1 lists the conditional power and type I errors given $Y^{(k)}$, with $\bar{Y}^{(k)}$ changing between two standard errors below and above θ_0 . For example, under n_0 , when the observed HC effect ($\bar{Y}^{(k)}$) is one standard error away ($\pm \sqrt{\sigma_0^2/m}$) from the true effect (θ_0), the mean power changes from 0.313 to 0.987 and the mean type I error from 0 to 0.088, which deviate far from the nominal levels of $1 - \beta = 0.8$ and $\alpha = 0.05$. Note that such a deviation is not a rare event because for a particular HC data set, there is a 32% chance that the sample mean is one standard error or further away from its true mean. The second and third columns of Figure 1 show that the distributions of type I error and power are extremely skewed, which is also revealed by the difference between their means and medians (achieved at $\bar{Y}^{(k)} = \theta_0$) in Table 1.

We briefly explain why the power and type I error have skewed distributions over the random realizations of the HC data. Taking the one-sample approach (n_1) for example, for a particular HC data $Y^{(k)}$, it can be shown that the conditional type I error is

$$h_1^{(k)} = 1 - \Phi \left(z_{1-\alpha} - \frac{\theta_0 - \bar{Y}^{(k)}}{\sqrt{\sigma_1^2/n_1}} \right).$$

In the parentheses, $z_{1-\alpha}$ is usually the dominant term and shifts the probability computation to the tail area of the normal distribution. As a result, although the sample mean ($\bar{Y}^{(k)}$) is symmetric around the true mean (θ_0), the impact of the sample mean being greater or smaller than the true mean is different. For $\bar{Y}^{(k)} > \theta_0$, the conditional type I errors have a range of $(0, \alpha)$, which is narrow under commonly specified significance levels. On the other hand, for $\bar{Y}^{(k)} < \theta_0$, the conditional type I errors have a much wider range, $(\alpha, 1)$. In summary, it is because researchers usually set power and significance level in the tail area (i.e., α close to 0 and $1 - \beta$ close to 1) that the random power and type I error are skewly distributed.

Finally, Table 1 provides empirical evidence for the theory presented in Theorem 2, which states a unified framework for n_v ($v = 0, 1, 2$).

Theorem 2. The sample sizes (n_0, n_1, n_2) control the random power and type I error in such a way that

1. The M-S approach (n_0) controls the mean of type I error at α and the median of power at $1 - \beta$;
2. The one-sample approach (n_1) controls the medians of type I error and power at α and $1 - \beta$, respectively;

3. The RCT approach (n_2) controls the means of type I error and power at α and $1 - \beta$, respectively.

Proof. See Appendix A.3.

Theorem 2 suggests that the M-S approach tries to reach a compromise between the one-sample and RCT approaches by controlling the mean type I error at α , while the median power at $1 - \beta$.

3 Sample Size Controlling the Percentiles

Simulation 1 shows that the distributions of power and type I error, observed over all the random realizations of HC data, are extremely skewed. For random variables with extremely skewed distributions, making decisions based on a location parameter such as a percentile is usually more desirable than based on the mean. We propose a sample size formula to control arbitrary percentiles of the random power and type I error. It provides a more sensible way to assess the risk in HCTs.

Theorem 3. Suppose in an HCT the goal is to control the $(1 - p_q)$ th percentile of the power at $1 - \beta$, and the p_h th percentile of the type I error at α . Then the required sample size is

$$n^* = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma_1^2}{\left[\Delta - (z_{p_q} + z_{p_h}) \sqrt{\sigma_0^2/m} \right]^2}, \quad (6)$$

and the null hypothesis is rejected if

$$Z^*(X, Y) = \frac{\bar{X} - \bar{Y} - z_{p_h} \sqrt{\sigma_0^2/m}}{\sqrt{\sigma_1^2/n^*}} > z_{1-\alpha}.$$

The parameters p_q and p_h can be specified arbitrarily as long as the condition

$$\Delta > (z_{p_q} + z_{p_h}) \sqrt{\sigma_0^2/m} \text{ holds.}$$

Proof. See Appendix A.4.

According to Theorem 3, let q^* and h^* be the random power and type I error under sample size n^* . Then we have $P(q^* > 1 - \beta) = p_q$ and $P(h^* < \alpha) = p_h$. Suppose an HCT is conducted to assess the effectiveness of a new drug. Given a particular HC data set, the realized power and type I error depend on the random difference between the HC sample mean ($\bar{Y}$) and its true mean (θ_0). However, if the researchers enroll n^* subjects, they can be confident that, over all the possible HC realizations, the power of the clinical trial would be greater than $1 - \beta$ with probability p_q , and the type I error would be smaller than α with probability p_h . It is easy to check that the sample size under the one-sample approach (n_1) is a special case of n^* at $p_q = p_h = 0.5$.

In other words, we propose sample size n^* to achieve the goal that, the operational characteristics (realized power and type I error) of an HCT are more desirable than the nominal levels with certain pre-specified probabilities (p_q and p_h). Controlling the power and type I error by percentiles instead of means (the RCT approach) is more effective when

their distributions are extremely skewed. Furthermore, the arbitrarily specified p_q and p_h provide flexibility in accommodating researchers' preference for risk control.

Based on the same setting as in Simulation 1, we conduct Simulation 2 to explore the properties of n^* . We consider different combinations of p_q and p_h , ranging from 0.5 to 0.9. Table 2 lists the estimated sample size n^* , the empirical means of type I error and power,

$$\bar{h}^* = \sum_{k=1}^K h^{*(k)} / K \text{ and } \bar{q}^* = \sum_{k=1}^K q^{*(k)} / K, \text{ and the empirical percentiles,}$$

$\widehat{p}_h = \sum_{k=1}^K I(h^{*(k)} < \alpha) / K$ and $\widehat{p}_q = \sum_{k=1}^K I(q^{*(k)} > 1 - \beta) / K$. From Table 2 we have two observations. First, for sample size calculation, p_q and p_h are exchangeable in the sense that switching their values leads to the identical sample size, which is also clear from Equation (6). Second, when the distributions of type I error and power are extremely skewed, it is more sensible to control the percentiles instead of means. For example, under $(p_h = 0.8, p_q = 0.7)$, we are confident that by enrolling 286 patients, the type I error would be smaller than 0.05 with probability 0.8, and the power would be greater than 0.8 with probability 0.7. It provides a high assurance for researchers. Note that in this case the mean power is 0.8 but the mean type I error is 0.06, slightly higher than the nominal $\alpha = 0.05$. As demonstrated in Table 1, the required sample size is $n_2 = 487$ (under the RCT approach) if we control the mean power and mean type I error at 0.8 and 0.05, respectively. That is, in order to reduce the mean type I error from 0.06 to 0.05, researchers need to enroll 201 additional patients, due to the skewness in the type I error distribution.

In simulations 1 and 2, we have assumed the population variances of the HC and experimental groups (σ_0^2 and σ_1^2) to be known, which is usually unrealistic in practice. We conduct Simulation 3 to further assess the performance of n^* in a more realistic scenario. It proceeds as follows: a) To compute the required sample size n^* , the assumed Δ and σ_1^2 will be plugged into Equation (6). However, σ_0^2 is replaced by s_0^2 , the HC sample variance. b) In hypothesis testing, we compute the test statistic $Z^*(X, Y)$ with σ_0^2 replaced by s_0^2 and σ_1^2 replaced by s_1^2 , the sample variance in the experimental group. Thus the estimated sample size includes additional uncertainty from s_0^2 , and the test statistic includes additional uncertainty from s_0^2 and s_1^2 . The detailed algorithm of Simulation 3 is presented in Appendix A.5.

Table 3 lists the results of Simulation 3. The sample size n^* becomes random when we replace the HC population variance (σ_0^2) with a random sample variance (s_0^2). When $p_q = p_h = 0.5$, the impact of the additional randomness is negligible. That is, after applying the integer restriction on sample size, the calculated n^* is constant at 69, the same as its counterpart in Table 2. As p_q or p_h increases, the standard deviation of n^* increases, and the mean of n^* deviates from the fixed sample size (in Table 2) computed under the population variance. This is because under large p_q and p_h , the distribution of n^* is heavily skewed to the right. Such skewness also arises from the tail behavior of the normal distribution. For example, under $(p_q = 0.9, p_h = 0.8)$, the random sample size has a mean of 3065.59 and a standard deviation of 29348.36. Its distribution has an extremely long tail (the 99th percentile is greater than 200000). The skewness is much more severe under $(p_q = p_h = 0.9)$, where we omit the simulation due to computer overflow. On the other hand, the operational characteristics of the clinical trial remain unchanged with additional randomness from the sample variances. Specifically, the realized controlling percentiles, $\widehat{p}_h$ and $\widehat{p}_q$, agree with the nominal levels. Furthermore, the means of power and type I error are close to those (in Table 2) obtained under the population variances. Taken together, the proposed sample size

n^* successfully controls the percentiles of power and type I error even when the population variances are unknown.

4 Example

The safety and efficacy of laparoscopic rectopexy for rectal prolapse will be compared with those of open rectopexy procedure, which was conducted several months ago by the same group of surgeons at the same institution [16]. Data will be collected prospectively for the laparoscopic rectopexy group and by hospital chart review for the HC group. The HC group includes 24 consecutive patients who had undergone conventional open rectopexy without having concomitant gynecologic procedures. These patients required an average of 71.5 milligrams of morphine during the first 48 hours after procedure with a standard deviation of 45.9 milligrams. It is expected that the average amount of morphine needed during the first 48 hours of laparoscopic rectopexy will be 41.5 milligrams with a standard deviation of 35.0 milligrams. We estimate the number of patients needed to detect the difference in morphine requirement during the first 48 hours between open and laparoscopic procedures, controlling the 70th percentile ($p_h = 0.7$) of type I error at 5%, and the 30th percentile ($p_q = 0.7$) of power at 80%. The numbers of patients needed in the laparoscopic group are $n_0 = 14$ (the M-S approach), $n_1 = 9$ (the one-sample approach), $n_2 = 22$ (the RCT approach), and $n^* = 19$ (the proposed approach), respectively. Note that the above sample sizes are obtained by replacing the unknown HC population variance with the sample variance.

5 Discussion

We have provided a unified framework for three existing approaches (M-S, one-sample, and RCT) in HCTs, by showing that they either control the mean or the median of power and type I error. We further developed a closed-form sample size formula to control arbitrary percentiles of the random power and type I error. It provides more flexibility in assessing the risk in HCTs and accommodates the extreme skewness in the distributions of power and type I error. We limited our discussion to the HCTs with continuous outcomes. In the future we will extend it to HCTs with binary and survival time outcomes.

Similar to the existing approaches, the proposed sample size formula (n^*) requires the population variances of the HC and the experimental group to be known. Through simulation study, we demonstrated that the proposed approach successfully controls the percentiles of power and type I error in a more realistic scenario, where the true variances are unknown and they are replaced with observed sample variances. One reviewer kindly pointed out that in situations where the measurements are continuous with bounded support, say on (a, b) , a sample size formula can be derived without requiring the population

variances. Specifically, we can define $\widehat{P} = (\bar{X} - a) / (b - a)$ with $0 < \widehat{P} < 1$. Applying the arcsin transformation on $\widehat{P}$, we can calculate sample size based on $\sin^{-1}(\sqrt{\widehat{P}})$, whose variance is free of the sampled population's true variance.

Lee and Tseng [14] presented sample size calculation for HCTs with binary outcomes controlling the means of power and type I error. Theorem 2 states that the same goal is achieved by n_2 for HCTs with continuous outcomes. The computation in [14] is more complicated due to the transformation performed on binary data. For continuous outcomes, when the HC variance is assumed to be known, the sample size formula does not depend on observations from the HC group. Thus one pair of null and alternative hypotheses leads to one unique sample size estimate. For binary outcomes, the sample size formula computed under the arcsin transformation depends on the observations from the HC group. Thus one pair of hypotheses leads to many possible sample size estimates, each determined by a

random realization of the HC data. In [14], the authors had to deal with the expectation of sample sizes.

The term $(\bar{Y} + z_{p_h} \sqrt{\sigma_0^2/m})$ in the numerator of $Z^*(\mathbf{X}, \mathbf{Y})$ is the p_h th percentile of posterior distribution $[\theta_0 | \mathbf{Y}]$ under a flat prior, which suggests a potential connection of the proposed approach to a Bayesian sample size calculation. Nonetheless, in Appendix A.4., the derivation of n^* is strictly in the frequentist paradigm, where the randomness of type I error and power comes from the uncertainty in the HC data $\mathbf{Y}$, not from random variable θ_0 (as in a Bayesian method). For example, we set type I error $h^* = \alpha$ at $\bar{Y} = \theta_0 - z_{p_h} \sqrt{\sigma_0^2/m}$, the $(1 - p_h)$ th percentile of $\bar{Y} | N(\theta_0, \sigma_0^2/m)$. Because h^* is monotonically decreasing in $\bar{Y}$, the p_h th percentile of h^* is controlled at α .

Acknowledgments

This study is supported in part by NIH grants UL1 RR024982 and P50 CA70907. The authors thank the two reviewers and associate editor for their constructive comments and suggestions.

Appendix A.1. Proof of Theorem 1

Proof. Assuming $\theta_0 - \bar{Y} = 0$ and applying some simple algebra, we transform (4) to

$$z_{1-\alpha} \sqrt{1 + n \frac{\sigma_0^2}{m\sigma_1^2}} = \Delta \sqrt{\frac{n}{\sigma_1^2}} - z_{1-\beta}.$$

Squaring on both sides and rearranging, we have

$$\left(\frac{z_{1-\alpha}^2 \sigma_0^2}{m\sigma_1^2} - \frac{\Delta^2}{\sigma_1^2} \right) n + 2 \frac{\Delta z_{1-\beta}}{\sqrt{\sigma_1^2}} \sqrt{n} + (z_{1-\alpha}^2 - z_{1-\beta}^2) = 0. \quad (7)$$

From (7) we can find a closed form solution for $\sqrt{n}$ subject to constraint that $\sqrt{n} > 0$.

First we need $b^2 - 4ac \geq 0$, where a , b and c are defined in (5). This condition implies

$\Delta \geq z_{1-\alpha} \sqrt{\sigma_0^2/m} \sqrt{1 - z_{1-\beta}^2/z_{1-\alpha}^2}$, and two possible roots

$$r_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \text{ and } \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

Fact 1. No plausible solution exists under $z_{1-\alpha} \sqrt{\sigma_0^2/m} \sqrt{1 - z_{1-\beta}^2/z_{1-\alpha}^2} \leq \Delta \leq z_{1-\alpha} \sqrt{\sigma_0^2/m}$.

- If $\Delta = z_{1-\alpha} \sqrt{\sigma_0^2/m}$ then $a = 0$, and the solution to (7) is $r = -c/b$. Because $c > 0$ when $\alpha < \beta$ and $b > 0$ by definition, we eliminate r due to the positive constraint on $\sqrt{n}$.
- If $z_{1-\alpha} \sqrt{\sigma_0^2/m} \sqrt{1 - z_{1-\beta}^2/z_{1-\alpha}^2} \leq \Delta < z_{1-\alpha} \sqrt{\sigma_0^2/m}$ then $a > 0$ and $4ac > 0$. It is easy to show that $r_1 < 0$ and $r_2 < 0$.

Sufficiency: We demonstrate that the condition $\Delta > z_{1-\alpha} \sqrt{\sigma_0^2/m}$ implies (5) being the unique sample size solution. From the condition we have $a < 0$ and $4ac < 0$. Together with $b > 0$, it is easy to show that $r_1 < 0$ and $r_2 > 0$. Thus (5) is the unique sample size solution.

Necessity: We demonstrate that (5) being the unique sample size solution implies

$\Delta > z_{1-\alpha} \sqrt{\sigma_0^2/m}$. (5) being the unique solution is equivalent to r_2 being the unique solution for $\sqrt{n}$. There are two scenarios:

1. $b^2 - 4ac = 0$ or $z_{1-\alpha} \sqrt{\sigma_0^2/m} \sqrt{1 - z_{1-\beta}^2/z_{1-\alpha}^2}$. It is eliminated due to Fact 1.
2. $b^2 - 4ac > 0$ and $r_1 < 0$ and $r_2 > 0$. Note that the condition $b^2 - 4ac > 0$ implies $\Delta > z_{1-\alpha} \sqrt{\sigma_0^2/m} \sqrt{1 - z_{1-\beta}^2/z_{1-\alpha}^2}$. We eliminate $z_{1-\alpha} \sqrt{\sigma_0^2/m} \sqrt{1 - z_{1-\beta}^2/z_{1-\alpha}^2} < \Delta \leq z_{1-\alpha} \sqrt{\sigma_0^2/m}$ based on Fact 1. The validity of $\Delta > z_{1-\alpha} \sqrt{\sigma_0^2/m}$ is established by Sufficiency.

Thus we complete the proof.

Appendix A.2. Algorithm of Simulation 1

Simulation 1. First we compute sample sizes n_v for a given set of $(m, \sigma_0^2, \sigma_1^2, \Delta, \alpha, \beta)$, where $v = 0, 1, 2$ denote the M-S, one-sample, and RCT approach, respectively. Then we generate null experimental data sets $\mathbf{X}_v^{0(l)} = \{X_1^{0(l)}, \dots, X_{n_v}^{0(l)}\}$ from $N(\theta_0, \sigma_1^2)$, and alternative experimental data sets $\mathbf{X}_v^{(l)} = \{X_1^{(l)}, \dots, X_{n_v}^{(l)}\}$ from $N(\theta_0 + \Delta, \sigma_1^2)$, for $l = 1, \dots, L$ and $v = 0, 1, 2$. The superscript 0 indicates that the null distribution is true, and the superscript(l) indicates the l th experimental data set generated. For iteration $k = 1, \dots, K$,

1. Simulate HC data $\mathbf{Y}^{(k)} = \{Y_1^{(k)}, \dots, Y_m^{(k)}\}$ from $N(\theta_0, \sigma_0^2)$;
2. Estimate the conditional type I error given $\mathbf{Y}^{(k)}$ by $h_v^{(k)} \approx \sum_{l=1}^L I(Z_v(\mathbf{X}_v^{0(l)}, \mathbf{Y}^{(k)}) > z_{1-\alpha}) / L$. Note that $Z_v(\mathbf{X}, \mathbf{Y}) = Z(\mathbf{X}, \mathbf{Y})$ for $v = 0$ and 2.
3. Estimate the conditional power given $\mathbf{Y}^{(k)}$ by $q_v^{(k)} \approx \sum_{l=1}^L I(Z_v(\mathbf{X}_v^{(l)}, \mathbf{Y}^{(k)}) > z_{1-\alpha}) / L$.

The superscript(k) of $h_v^{(k)}$ and $q_v^{(k)}$ suggests that they are computed given the k th simulated HC data. We set $K = L = 5000$.

Appendix A.3. Proof of Theorem 2

Proof. We first state the fact that $\bar{Y} \sim N(\theta_0, \sigma_0^2/m)$, $\bar{X}_v \sim N(\theta_0 + \Delta, \sigma_1^2/n_v)$, $\bar{X}_v^0 \sim N(\theta_0, \sigma_1^2/n_v)$, and $\text{median}(\bar{Y}) = \theta_0$. For n_0 and n_2 , the null hypothesis is rejected if $Z(\mathbf{X}, \mathbf{Y}) > z_{1-\alpha}$. Marginalizing with respect to $\mathbf{Y}$ is equivalent to marginalizing with respect to $\bar{Y}$. Thus for $v = 0, 2$,

$$\begin{aligned}
E(h_v) &= \int P\left(\frac{\bar{X}_v - \bar{Y}}{\sqrt{\sigma_1^2/n_v + \sigma_0^2/m}} > z_{1-\alpha} \mid \bar{Y}\right) f(\bar{Y}) d\bar{Y} \\
&= \int I\left(\frac{\bar{X}_v - \bar{Y}}{\sqrt{\sigma_1^2/n_v + \sigma_0^2/m}} > z_{1-\alpha}\right) f(\bar{X}_v) f(\bar{Y}) d\bar{X}_v d\bar{Y} \\
&= \int_{z_{1-\alpha}}^{+\infty} f(U_v) dU_v = P(U_v > z_{1-\alpha}) = \alpha.
\end{aligned}$$

We have the third equality through random variable transformation, where

$$U_v = \left(\bar{X}_v - \bar{Y}\right) / \sqrt{\sigma_1^2/n_v + \sigma_0^2/m} \text{ and it is easy to show that } U_v \sim N(0, 1).$$

In the similar fashion, we can show that $E(q_2) = 1 - \beta$,

$$\begin{aligned}
E(q_2) &= \int P\left(\frac{\bar{X}_2 - \bar{Y}}{\sqrt{\sigma_1^2/n_2 + \sigma_0^2/m}} > z_{1-\alpha} \mid \bar{Y}\right) f(\bar{Y}) d\bar{Y} \\
&= P\left(U > z_{1-\alpha} - \frac{\Delta}{\sqrt{\sigma_1^2/n_2 + \sigma_0^2/m}}\right) = 1 - \beta.
\end{aligned}$$

We have the second equality by defining $U = (\bar{X} - \bar{Y} - \Delta) / \sqrt{\sigma_1^2/n_v + \sigma_0^2/m}$. The third equality is obtained by plugging the expression of n_2 and recognizing $U \sim N(0, 1)$

We then show that $\text{median}(q_0) = 1 - \beta$. From (3) we have

$$q_0^{(k)} = 1 - \Phi\left(\frac{z_{1-\alpha} \sqrt{\sigma_1^2/n_0 + \sigma_0^2/m}}{\sqrt{\sigma_1^2/n_0}} - \frac{\Delta + \theta_0 - \bar{Y}^{(k)}}{\sqrt{\sigma_1^2/n_0}}\right).$$

First we recognize that $q_0^{(k)}$ is a decreasing function of $\bar{Y}^{(k)}$. Second, n_0 is the solution to $q_0^{(k)} = 1 - \beta$ by setting $\bar{Y}^{(k)} = \theta_0 = \text{median}(\bar{Y})$. These two points lead to the conclusion that $\text{median}(q_0) = 1 - \beta$. Note that $E(q_2)$ and $q_0^{(k)}$ have different expressions because the former marginalizes with respect to random $\bar{Y}$, while the latter is defined conditional on a particular $\bar{Y}^{(k)}$.

Now we show that $\text{median}(h_1) = \alpha$,

$$h_1^{(k)} = P\left(\frac{\bar{X}_0 - \bar{Y}^{(k)}}{\sqrt{\sigma_1^2/n_1}} > z_{1-\alpha}\right) = 1 - \Phi\left(z_{1-\alpha} - \frac{\theta_0 - \bar{Y}^{(k)}}{\sqrt{\sigma_1^2/n_1}}\right).$$

Thus $h_1^{(k)}$ is a decreasing function of $\bar{Y}^{(k)}$ and $h_1^{(k)} = \alpha$ at $\bar{Y}^{(k)} = \theta_0 = \text{median}(\bar{Y})$. Thus we conclude $\text{median}(h_1) = \alpha$. Similar argument leads to the conclusion that $\text{median}(q_1) = 1 - \beta$.

Appendix A.4. Proof of Theorem 3

Proof. First we demonstrate that based on $Z^*(X, Y)$, the p_h th percentile of type I error is controlled at α : For a given $\bar{Y}$, the type I error is

$$h^* = P\left(\frac{\bar{X} - \bar{Y} - z_{p_h} \sqrt{\sigma_0^2/m}}{\sqrt{\sigma_1^2/n^*}} > z_{1-\alpha}\right) = 1 - \Phi\left(z_{1-\alpha} + \frac{\bar{Y} - (\theta_0 - z_{p_h} \sqrt{\sigma_0^2/m})}{\sqrt{\sigma_1^2/n^*}}\right).$$

Thus $h^* = \alpha$ when $\bar{Y} = \theta_0 - z_{p_h} \sqrt{\sigma_0^2/m}$, which is the $(1-p_h)$ th percentile of $\bar{Y} \sim N(\theta_0, \sigma_0^2/m)$.

Together with the fact that h^* is a monotonically decreasing function in $\bar{Y}$, we have

$$P(h^* < \alpha) = P(\bar{Y} > \theta_0 - z_{p_h} \sqrt{\sigma_0^2/m}) = p_h. \text{ Note that this statement holds for any } n^*.$$

Then we solve for n^* which controls the $(1-p_q)$ th percentile of power at $1-\beta$: The conditional power given $\bar{Y}$ is

$$q^* = P\left(\frac{\bar{X} - \bar{Y} - z_{p_h} \sqrt{\sigma_0^2/m}}{\sqrt{\sigma_1^2/n^*}} > z_{1-\alpha}\right) = 1 - \Phi\left(z_{1-\alpha} + \frac{\bar{Y} - \theta_0 - \Delta + z_{p_h} \sqrt{\sigma_0^2/m}}{\sqrt{\sigma_1^2/n^*}}\right). \quad (8)$$

It is obvious that q^* is a monotonically decreasing with $\bar{Y}$. Using this property, if we set $q^* =$

$1-\beta$ at $\bar{Y} = \theta_0 + z_{p_q} \sqrt{\sigma_0^2/m}$, the p_q th percentile of $\bar{Y}$, we can achieve the goal of controlling the $(1-p_q)$ th percentile of power at $1-\beta$, because

$$P(q^* > 1-\beta) = P(\bar{Y} < \theta_0 + z_{p_q} \sqrt{\sigma_0^2/m}) = p_q.$$

Thus by plugging $q^* = 1-\beta$ and $\bar{Y} = \theta_0 + z_{p_q} \sqrt{\sigma_0^2/m}$ into (8), we can solve for n^* from the following equation,

$$1-\beta = 1 - \Phi\left(z_{1-\alpha} + \frac{(\theta_0 + z_{p_q} \sqrt{\sigma_0^2/m}) - \theta_0 - \Delta + z_{p_h} \sqrt{\sigma_0^2/m}}{\sqrt{\sigma_1^2/n^*}}\right).$$

The solution for n^* , equation (6), can be obtained after some algebra. The condition

$\Delta > (z_{p_q} + z_{p_h}) \sqrt{\sigma_0^2/m}$ is due to the positive constraint on $\sqrt{n^*}$.

Appendix A.5. Algorithm of Simulation 3

Simulation 3. For iteration $k = 1, \dots, K$,

1. Generate HC data $\mathbf{Y}^{(k)} = \{Y_1^{(k)}, \dots, Y_m^{(k)}\}$ from $N(\theta_0, \sigma_0^2)$. Compute the sample variance $s_0^{2(k)}$;

2. Estimate the required sample size $n^{*(k)}$ based on Formula (6), with σ_0^2 replaced by $s_0^{2(k)}$;
3. Given sample size $n^{*(k)}$, generate null experimental data sets $\mathbf{X}^{0(k,l)} = \{X_1^{0(l)}, \dots, X_{n^{*(k)}}^{0(l)}\}$ from $N(\theta_0, \sigma_1^2)$, and alternative experimental data sets $\mathbf{X}^{(k,l)} = \{X_1^{(l)}, \dots, X_{n^{*(k)}}^{(l)}\}$ from $N(\theta_0 + \Delta, \sigma_1^2)$, for $l = 1, \dots, L$;
4. Compute the empirical type I error by $h^{*(k)} \approx \sum_{l=1}^L I(Z^*(\mathbf{X}^{0(k,l)}, \mathbf{Y}^{(k)}) > z_{1-\alpha}) / L$. Note that we replace the population variances (σ_0^2 and σ_1^2) in $Z^*(\mathbf{X}^{0(k,l)}, \mathbf{Y}^{(k)})$ by sample variances ($s_0^{2(k)}$ and $s_1^{2(k,l)}$). Here $s_1^{2(k,l)}$ is the sample variance of $\mathbf{X}^{0(k,l)}$. Similarly, we compute the empirical power $q^{*(k)}$.

References

- [1]. Makuch RW, Simon RM. Sample size considerations for non-randomized comparative studies. *Journal of Chronic Diseases*. 1980; 33(3):175–181. [PubMed: 7354106]
- [2]. Vickers AJ, Ballen V, Scher HI. Setting the bar in phase II trials: The use of historical data for determining "go/no go" decision for definitive phase III testing. *Clinical Cancer Research*. 2007; 13(3):972–976. [PubMed: 17277252]
- [3]. Cho SD, Krishnaswami S, Mckee JC, Zallen G, Silen ML, Bliss DW. Analysis of 29 consecutive thoracoscopic repairs of congenital diaphragmatic hernia in neonates compared to historical controls. *Journal of Pediatric Surgery*. 2009; 44(1):80–86. [PubMed: 19159722]
- [4]. Abe T, Kakemura T, Fujinuma S, Maetani I. Successful outcomes of emr-l with 3d-eus for rectal carcinoids compared with historical controls. *World Journal of Gastroenterology*. 2008; 14(25):4054–4058. [PubMed: 18609690]
- [5]. Storm C, Steffen I, Schefold JC, Krueger A, Oppert M, Jorres A, Hasper D. Mild therapeutic hypothermia shortens intensive care unit stay of survivors after out-of-hospital cardiac arrest compared to historical controls. *Critical Care*. 12(3):2008.
- [6]. Van Rooij WJ, De Gast AN, Sluzewski M. Results of 101 aneurysms treated with polyglycolic/polylactic acid microfilament nexus coils compared with historical controls treated with standard coils. *American Journal of Neuroradiology*. 2008; 29(5):991–996. [PubMed: 18310232]
- [7]. Ando R, Nakamura A, Nagatani M, Yamakawa S, Ohira T, Takagi M, Matsushima K, Aoki A, Fujita Y, Tamura K. Comparison of past and recent historical control data in relation to spontaneous tumors during carcinogenicity testing in fischer 344 rats. *Journal of Toxicologic Pathology*. 2008; 21(1):53–60.
- [8]. Song JY, Chung BS, Choi KC, Shin BS. A 5-year period clinical observation on herpes zoster and the incidence of postherpetic neuralgia (2002–2006); a comparative analysis with the historical control group of a previous study (1995–1999). *Korean Journal of Dermatology*. 2008; 46(4):431–436.
- [9]. Loudon I. The use of historical controls and concurrent controls to assess the effects of sulphonamides, 1936–1945. *Journal of the Royal Society of Medicine*. 2008; 101(3):148–155. [PubMed: 18344473]
- [10]. Dixon DO, Simon R. Sample size considerations for studies comparing survival curves using historical controls. *Journal of Clinical Epidemiology*. 1988; 41(12):1209–1213. [PubMed: 3062140]
- [11]. Chang MN, Shuster JJ, Kepner JL. Group sequential designs for phase II trials with historical controls. *Controlled Clinical Trials*. 1999; 20(4):353–364. [PubMed: 10440562]
- [12]. Kepner J, Wackerly D. Some observations on the makuch/simon approach to sample size determination in clinical trials with historical controls. *Communications in Statistics Part B: Simulation and Computation*. 2001; 30(3):611–621.

- [13]. Chang MN, Shuster JJ, Kepner JL. Sample sizes based on exact unconditional tests for phase II clinical trials with historical controls. *Journal of Biopharmaceutical Statistics*. 2004; 14(1):189–200. [PubMed: 15027508]
- [14]. Lee JJ, Tseng C. Uniform power method for sample size calculation in historical control studies with binary response. *Controlled Clinical Trials*. 2001; 22(4):390–400. [PubMed: 11514040]
- [15]. Korn EL, Freidlin B. Conditional power calculations for clinical trials with historical controls. *Statistics in Medicine*. 2006; 25(17):2922–2931. [PubMed: 16479548]
- [16]. Abraham NS, DuraiRaj R, Young JM, Young CJ, Solomon MJ. How does an historic control study of a surgical procedure compare with the "gold standard"? *Diseases of the Colon and Rectum*. 2006; 49(8):1141–1148. [PubMed: 16841269]

Figure 1.

The type I errors ($h_v^{(k)}$) and powers ($q_v^{(k)}$) under n_0 , n_1 , and n_2 . The first column plots $h_v^{(k)}$ and $q_v^{(k)}$ versus the difference between the observed and true HC effects, $\bar{Y}^{(k)} - \theta_0$, with black for $h_v^{(k)}$ and red for $q_v^{(k)}$. The second and third columns plot the histograms of h_v and q_v , respectively (the vertical axes indicate probability density).

Table 1

Simulation 1, Conditional Type I Errors and Powers Given $\mathbf{Y}^{(k)}$

HC sample mean $\bar{Y}^{(k)}$	$n_0 = 144$		$n_1 = 69$		$n_2 = 487$	
	$h_0^{(k)}$	$q_0^{(k)}$	$h_1^{(k)}$	$q_1^{(k)}$	$h_2^{(k)}$	$q_2^{(k)}$
2 standard errors below θ_0	0.482	0.999	0.597	0.996	0.733	1.000
1 standard error below θ_0	0.088	0.987	0.248	0.963	0.025	1.000
θ_0	0.003	0.801	0.054	0.822	0.000	0.988
1 standard error above θ_0	0.000	0.313	0.004	0.482	0.000	0.418
2 standard errors above θ_0	0.000	0.032	0.000	0.162	0.000	0.003
Marginal Mean	0.050	0.695	0.115	0.742	0.048	0.802

n_0 , n_1 and n_2 are sample sizes under the Makuch-Simon (M-S), one-sample, and RCT approaches. θ_0 is the true effect in the historical control (HC) group. $h_V^{(k)}$ and $q_V^{(k)}$ indicate the conditional type I error and power given historical control data $\mathbf{Y}^{(k)}$.

Table 2

Simulation 2, Type I Errors and Powers under n^*

	p_q				
	0.5	0.6	0.7	0.8	0.9
0.5	69 (0.11, 0.73) [0.51, 0.51]	84 (0.12, 0.78) [0.49, 0.61]	107 (0.13, 0.83) [0.51, 0.69]	146 (0.16, 0.88) [0.49, 0.79]	252 (0.20, 0.94) [0.49, 0.90]
0.6	84 (0.09, 0.72) [0.58, 0.49]	105 (0.10, 0.78) [0.61, 0.61]	137 (0.11, 0.83) [0.62, 0.70]	197 (0.13, 0.88) [0.61, 0.80]	376 (0.17, 0.94) [0.60, 0.90]
0.7	107 (0.07, 0.72) [0.70, 0.52]	137 (0.08, 0.76) [0.69, 0.60]	186 (0.08, 0.82) [0.71, 0.71]	286 (0.10, 0.87) [0.70, 0.80]	643 (0.14, 0.93) [0.71, 0.90]
0.8	146 (0.05, 0.69) [0.81, 0.50]	197 (0.05, 0.75) [0.81, 0.60]	286 (0.06, 0.80) [0.82, 0.70]	495 (0.08, 0.86) [0.82, 0.80]	1577 (0.11, 0.93) [0.82, 0.90]
0.9	252 (0.03, 0.66) [0.91, 0.49]	376 (0.03, 0.72) [0.90, 0.60]	643 (0.04, 0.78) [0.91, 0.70]	1577 (0.05, 0.84) [0.90, 0.80]	34246 (0.08, 0.91) [0.90, 0.90]

The estimated sample size n^* , the average empirical type I errors and empirical powers, $(\hat{h} \cdot \hat{q})$, and the realized controlling percentiles, $[\hat{p}_h, \hat{p}_q]$, under different combinations of p_h and p_q . We assume $m = 80$, $\sigma_0^2 = \sigma_1^2 = 1$, $\Delta = 0.3$, $\alpha = 0.05$, $1 - \beta = 0.8$, and the population variances (σ_0^2 and σ_1^2) are known.

Table 3

Simulation 3, Type I Errors and Powers under n^*

		p_q				
		0.5	0.6	0.7	0.8	0.9
p_h	0.5	69(0) (0.11, 0.74) [0.50, 0.51]	84.23(1.42) (0.12, 0.78) [0.50, 0.61]	106.56(4.08) (0.14, 0.83) [0.50, 0.70]	146.43(10.60) (0.16, 0.88) [0.48, 0.80]	254.63(37.63) (0.20, 0.94) [0.50, 0.90]
	0.6	84.23(1.42) (0.09, 0.72) [0.60, 0.51]	104.83(3.86) (0.10, 0.77) [0.60, 0.60]	136.75(8.84) (0.11, 0.82) [0.62, 0.70]	197.38(21.74) (0.13, 0.88) [0.60, 0.80]	385.16(85.93) (0.17, 0.94) [0.60, 0.90]
	0.7	106.56(4.08) (0.07, 0.71) [0.70, 0.50]	136.75(8.84) (0.07, 0.76) [0.71, 0.60]	186.25(19.06) (0.08, 0.81) [0.71, 0.70]	289.38(48.89) (0.10, 0.87) [0.71, 0.80]	691.34(262.40) (0.14, 0.93) [0.70, 0.89]
	0.8	146.43(10.60) (0.05, 0.69) [0.80, 0.50]	197.38(21.74) (0.05, 0.75) [0.81, 0.60]	289.38(48.89) (0.06, 0.80) [0.82, 0.70]	517.52(151.21) (0.08, 0.86) [0.81, 0.80]	3065.59(29348.36) (0.11, 0.92) [0.81, 0.89]
	0.9	254.63(37.63) (0.03, 0.66) [0.90, 0.49]	385.16(85.93) (0.03, 0.72) [0.90, 0.59]	691.34(262.40) (0.04, 0.78) [0.90, 0.70]	3065.59(29348.36) (0.05, 0.84) [0.90, 0.79]	- - -

The estimated mean(standard deviation) of n^* , the average empirical type I errors and empirical powers, $(\hat{p}_h, \hat{p}_q)$, and the realized controlling percentiles, $[\hat{p}_h, \hat{p}_q]$ under different combinations of p_h and p_q . We assume $m = 80$, $\sigma_0^2 = \sigma_1^2 = 1$, $\Delta = 0.3$, $\alpha = 0.05$, $1 - \beta = 0.8$, and the population variances (σ_0^2 and σ_1^2) are unknown.