Published in final edited form as:

J Biopharm Stat. 2016 ; 26(4): 657–671. doi:10.1080/10543406.2015.1052495.

Sample size considerations for historical control studies with survival outcomes

Hong Zhu*, Song Zhang, and Chul Ahn

Division of Biostatistics, Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas, USA

Abstract

Historical control trials (HCTs) are frequently conducted to compare an experimental treatment with a control treatment from a previous study, when they are applicable and favored over a randomized clinical trial (RCT) due to feasibility, ethics and cost concerns. Makuch and Simon developed a sample size formula for historical control (HC) studies with binary outcomes, assuming that the observed response rate in the HC group is the true response rate. This method was extended by Dixon and Simon to specify sample size for HC studies comparing survival outcomes. For HC studies with binary and continuous outcomes, many researchers have shown that the popular Makuch and Simon method does not preserve the nominal power and type I error, and suggested alternative approaches. For HC studies with survival outcomes, we reveal through simulation that the conditional power and type I error over all the random realizations of the HC data have highly skewed distributions. Therefore, the sampling variability of the HC data needs to be appropriately accounted for in determining sample size. A flexible sample size formula that controls arbitrary percentiles, instead of means, of the conditional power and type I error, is derived. Although an explicit sample size formula with survival outcomes is not available, the computation is straightforward. Simulations demonstrate that the proposed method preserves the operational characteristics in a more realistic scenario where the true hazard rate of the HC group is unknown. A real data application of an advanced non-small cell lung cancer (NSCLC) clinical trial is presented to illustrate sample size considerations for HC studies in comparison of survival outcomes.

Keywords

Clinical trial; Historical control; Percentiles of type I error and power; Sample size; Survival outcome

1 Introduction

Although randomized clinical trials (RCTs) are the gold standard for comparing different treatments, there are situations where historical control trials (HCTs) are favored over an RCT due to considerations of feasibility, ethics, economics, and regulatory acceptability. In

* Address correspondence to Dr. Hong Zhu, Division of Biostatistics, Department of Clinical Sciences, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, Texas 75390, USA; ; Email: hong.zhu@utsouthwestern.edu

an HCT, an experimental treatment is compared with a control treatment from a previously conducted trial. HCTs have a major advantage of avoiding randomizing patients into potentially inferior treatments. Another benefit of HCTs is the common perception that this type of single-arm, non-randomized trial requires a much smaller sample size and is easier to conduct than its RCT counterpart. Although historical control (HC) design is prone to patient selection bias, outcome evaluation bias and undefinitive conclusions, the use of HCTs provides at least a tentative comparison of the experimental treatment with the control. The HC design has long been discussed in statistical literature (Gehan, 1984; Pinsky, 1984; Keiding, 1995; Ibrahim et al., 1998), and has been widely applied in clinical practice (Loudon, 2008; Buchs and Silverman, 2011; Chen et al., 2012).

Here we briefly review the basic idea of existing sample size approaches for HC studies. In each particular HCT, the observed HC group is a realization from the underlying unknown HC distribution. Makuch and Simon (1980) derived a sample size formula using the same two-sample test statistic for proportions as in RCTs but only allowing the experimental group to be random. It requires the (unrealistic) assumption that the observed HC proportion of successes equals the true HC proportion of successes. As a result, many researchers (Kepner and Wackerly, 2001; Lee and Tseng, 2001; Korn and Freidlin, 2006) have demonstrated that this method could not preserve the nominal power and type I error when taking uncertainty about the HC data into account. To facilitate discussion about this uncertainty, we introduce the concepts of conditional power and conditional type I error for historical control studies with survival outcomes. For a given sample size, the power and type I error that are actually achieved depend on the randomly realized relationship between the observed HC hazard rate ($\hat{\lambda}_c$) and the true HC hazard rate (λ_c). Let λ_e denote the true experimental hazard rate. Basically, if $\hat{\lambda}_c$ happens to be greater than λ_c , it will be more likely to accept $H_a: \lambda_c > \lambda_e$, resulting in a greater power but also a greater type I error. If $\hat{\lambda}_c$ happens to be smaller than λ_c , however, it will be less likely to accept $H_a: \lambda_c > \lambda_e$, resulting in a smaller power and a smaller type I error. The above discussion suggests that the power and type I error of a particular HCT should be discussed in a “conditional” sense, given the randomly observed HC data. The uniform or unconditional power approaches (Lee and Tseng, 2001; Korn and Freidlin, 2006) calculate sample sizes so that the means of conditional powers and type I errors over all possible realizations of HC data equal to the desired levels. Such approaches take a hypothetical long-term view, assuming that the HCT will be performed infinite times, each time with a different random realization of HC data, and the averaged performance is considered important. In reality, however, an HCT is usually conducted only once based on one realization of HC data. Instead of the means of conditional powers and type I errors over hypothetical repetitions, a trial characteristic more relevant to researchers might be: what is the probability of actually achieving the desired power and type I error in this trial? Zhang et al. (2010) and (2013) suggested a different sample size approach such that the operational characteristics (conditional power and type I error) of an HCT will be more desirable than nominal levels with pre-specified probabilities. To be specific, they proposed to control the conditional power and type I error by their percentiles instead of means. This sample size approach is further justified by the observation that the distributions of conditional powers and type I errors are highly skewed.

It is a basic principle in statistics that for a highly skewed variable, making decision based on its percentile instead of mean is usually more meaningful.

Zhang et al. (2010) and (2013) developed sample size formulas that control the percentiles of the conditional power and type I error for HC studies with continuous and binary outcomes. In this study we propose to extend this approach to HC studies with survival outcomes. Many chronic disease studies, such as cancer and cardiovascular diseases, often have the time to an event, such as death or recurrence of a disease, as the primary endpoint. Cancer researchers usually perform a Phase II study of a new treatment to obtain an estimate of its anti-tumor effect on time to progression, prior to performing a comparative Phase III study. The use of an HC study could produce a reasonably powerful preliminary comparison with a control group, in addition to the estimate of anti-tumor activity. If such a study is to be carried out, it should be designed appropriately to ensure sufficient power. Despite an extensive literature on specification of sample size for HC studies with binary and continuous outcomes, only a few HC studies are available for survival outcomes (Dixon and Simon, 1988; Emrich, 1989). The comparison of survival outcomes is more complex than the comparison of binary or continuous outcomes because of the presence of censored observations, and the possibility of additional data collection of control group subjects who had not yet failed at the time of designing a new prospective study. The focus of this paper is to propose an alternative method of determining the required sample size, for studies comparing survival outcomes, when a control group from a previous trial is already available. Following the paradigm of Dixon and Simon (1988), we consider the comparison of an experimental treatment with a control treatment assuming exponentially distributed survival times and uniform censoring. The proposed method appropriately accounts for the random variability of the HC data in sample size determination.

We organize the remainder of the article as follows. In Section 2, standard methods of sample size calculation for HC studies with survival outcomes are reviewed. Their performances are assessed through simulation under the omniscient model, where the true hazard rates in the HC and experimental groups are known. It demonstrates highly skewed distributions of the conditional power and type I error over all the random realizations of the HC data. In Section 3, we present the idea of controlling arbitrary percentiles of the conditional power and type I error, and derive a flexible sample size formula, which can help clinicians to better understand how various design factors impact the sample size. The proposed method is evaluated by simulations in both an ideal scenario with true HC hazard rate known and a more realistic scenario with the true HC hazard rate unknown in Section 4. In Section 5, a real data application of a metastatic breast cancer trial is presented for illustration. We provide concluding remarks and discussion in Section 6.

2 Review of standard methods

For sample size considerations with survival outcomes, the following assumptions are made as specified in Dixon and Simon (1988).

1. Failure times are exponentially distributed with hazard λ_c in the control group and hazard λ_e in the experimental group.

2. Subjects are uniformly accrued during accrual period T and followed during follow-up period τ after completion of accrual, so the total duration of the study is $T + \tau$.
3. There is no drop-out and every subject is followed until the event or the end of study. Thus, censoring times follow a uniform(τ , $T + \tau$) distribution.

For one-sided two sample test to compare two exponential survival distributions, $H_0: \lambda_c = \lambda_e$ vs $H_a: \lambda_c > \lambda_e$, the test statistic is

$$W = \frac{\log(\hat{\lambda}_c) - \log(\hat{\lambda}_e)}{\sqrt{1/d_c + 1/d_e}},$$

where $\hat{\lambda}_c$ and d_c are estimated hazard and the number of observed failures in the control group, and $\hat{\lambda}_e$ and d_e are estimated hazard and the number of observed failures in the experimental group. The null hypothesis is rejected if $W > Z_{1-\alpha}$, where $Z_{1-\alpha}$ is the $100(1 - \alpha)$ -th percentile of the standard normal distribution. By standard asymptotic likelihood method, we have

$$\log \frac{\hat{\lambda}_c}{\hat{\lambda}_e} \sim N\left(\log \frac{\lambda_c}{\lambda_e}, \frac{1}{E(d_c)} + \frac{1}{E(d_e)}\right),$$

where $E(d_c)$ and $E(d_e)$ are the expected numbers of failures in the control and experimental groups, depending on λ_c , λ_e , the censoring distribution, and the accrual and follow-up periods. With exponential assumption, the hazard for the control group can be estimated by $\hat{\lambda}_c = d_c/x_c$, where x_c is the sum of both failure times and censoring times in the control group, and the hazard for the experimental group can be estimated by $\hat{\lambda}_e = d_e/x_e$, where x_e is the sum of both failure times and censoring times in the experimental group. Let $\delta = \lambda_c/\lambda_e$ denote the hazard ratio. The null and alternative hypotheses can then be re-written as $H_0: \delta = 1$ vs $H_a: \delta > 1$.

For the exponential survival and uniform censoring distribution, the expected number of failures in the experimental group is

$$E(d_e) = n_e \left[1 - \frac{\exp(-\tau\lambda_e)}{T\lambda_e} \{1 - \exp(-T\lambda_e)\} \right],$$

where n_e is the sample size of the experimental group (Dixon and Simon, 1988). Replacing n_e by $n_e = rT$, where r is accrual rate and T is accrual duration, then

$$E(d_e) = r[T - \exp(-\tau\lambda_e)\{1 - \exp(-T\lambda_e)\}/\lambda_e]. \quad (1)$$

On the other hand, with power $1 - \beta$, type I error α , true hazard rate λ_c in the control group and hazard ratio δ , the expected number of failures in the experimental group, $E(d_e)$, can be determined by the following power equation

$$P\left(\frac{\log(\hat{\lambda}_c) - \log(\hat{\lambda}_e)}{\sqrt{E(d_c)^{-1} + E(d_e)^{-1}}} > Z_{1-\alpha}\right) = 1 - \beta. \quad (2)$$

Apparently, the power depends on the censoring distribution, and the accrual and follow-up periods. For follow-up studies with survival outcomes, it is often desirable to compute required accrual duration T with specified accrual rate r and other design parameters. Combining equations (1) and (2), it is straightforward to find the required accrual duration T . The sample size needed for the experimental group is obtained as $n_e = rT$. Further, although the exponential assumption may seem restrictive for determining sample size, this approach is known to work well compared with a more general approach which adopts the proportional hazards assumption and uses a log-rank test (Korn and Freidlin, 2006).

2.1 Randomized trial method

Under the randomized trial method, the HC group is treated as a regular control arm in a randomized trial by including the variability of $\hat{\lambda}_c$. The power equation (2) can be re-written as

$$P\left(\frac{\log(\hat{\lambda}_c/\hat{\lambda}_e) - \log\delta}{\sqrt{E(d_e)^{-1} + E(d_c)^{-1}}} > Z_{1-\alpha} - \frac{\log\delta}{\sqrt{E(d_e)^{-1} + E(d_c)^{-1}}}\right) = 1 - \phi\left(Z_{1-\alpha} - \frac{\log\delta}{\sqrt{E(d_e)^{-1} + E(d_c)^{-1}}}\right) = 1 - \beta,$$

where ϕ is the cumulative distribution function of the standard normal distribution. The expected number of failures in the experimental group is thus given by

$$E(d_e) = \frac{E(d_c)(Z_{1-\alpha} + Z_{1-\beta})^2}{E(d_c)(\log\delta)^2 - (Z_{1-\alpha} + Z_{1-\beta})^2}. \quad (3)$$

From (1), we have $E(d_e) = r[T - \exp(-\tau\lambda_e)\{1 - \exp(-T\lambda_e)\}/\lambda_e]$ where $\lambda_e = \lambda_c/\delta$. For a given δ , replacing λ_c by $\hat{\lambda}_c$ and $E(d_c)$ by d_c , the accrual duration T and sample size $n_e = rT$ for the experimental group can be computed by equations (1) and (3). We use T_e^R and n_e^R to denote the accrual duration and sample size for the experimental group by the randomized trial method. Unlike the continuous outcomes for which sample size of the experimental group does not rely on specific values of an HC group, n_e^R for survival outcomes depends on $\hat{\lambda}_c$, the estimated HC hazard.

2.2 Dixon and Simon method

Under the Dixon and Simon (DS) method, the HC observations are assumed not subject to variability because they have already been observed. The power equation (2) becomes

$$P \left(\frac{\log(\hat{\lambda}_c/\hat{\lambda}_e) - \log(\hat{\lambda}_c/\lambda_e)}{\sqrt{E(d_e)^{-1}}} > \frac{Z_{1-\alpha} \sqrt{E(d_e)^{-1} + E(d_c)^{-1}}}{\sqrt{E(d_e)^{-1}}} - \frac{\log(\hat{\lambda}_c/\lambda_e)}{\sqrt{E(d_e)^{-1}}} \simeq 1 - \phi \left(\frac{Z_{1-\alpha} \sqrt{E(d_e)^{-1} + E(d_c)^{-1}}}{\sqrt{E(d_e)^{-1}}} - \frac{\log \delta}{\sqrt{E(d_e)^{-1}}} \right) = 1 - \beta,$$

where we use the approximation $\delta = \lambda_c/\lambda_e \simeq \hat{\lambda}_c/\hat{\lambda}_e$. The expected number of failures in the experimental group, $E(d_e)$, can be determined by the following equation

$$Z_{1-\alpha} \sqrt{E(d_e)^{-1} + E(d_c)^{-1}} + Z_{1-\beta} \sqrt{E(d_e)^{-1}} = \log \delta. \quad (4)$$

Similarly, combining equations (1) and (4), and replacing λ_c by $\hat{\lambda}_c$ and $E(d_c)$ by d_c , we can compute the accrual duration T and sample size $n_e = rT$ for the experimental group. We use T_e^D and n_e^D to denote the accrual duration and sample size for the experimental group by the DS method.

2.3 Performance evaluation

The performance of the randomized trial method and DS method for HC studies are evaluated under the omniscient model by simulation, where (n_c, d_c, δ) are given and the unknown true hazard λ_c is estimated by the observed HC data. The estimated hazard $\hat{\lambda}_c$ in the HC group is a random realization centered around true λ_c . The random variability in the HC data not only affects sample size required for the experimental group by both the randomized trial method and DS method, but also affects the test statistic W that depends on $\hat{\lambda}_c$ and d_c . Therefore, given a particular random realization of the HC data, we have conditional sample size for the experiment group, conditional type I error and conditional power. We conduct simulation studies to explore the operational characteristics (conditional type I error and power) of the HC study with survival outcomes over all the possible realizations.

The simulation procedure is as follows. Given $(\lambda_c, n_c, d_c, \delta, r, \tau)$ and for iteration $k = 1, \dots, K$,

1. Generate n_c survival times from exponential distribution with hazard λ_c and generate n_c censoring times from uniform distribution in the HC group to ensure d_c observed failures. Compute estimated hazard $\hat{\lambda}_c^{(k)} = d_c/x_c^{(k)}$ in the HC group. The superscript k indicates that the terms arise from the k -th HC data.
2. Compute the accrual durations, $T_e^{R(k)}$ and $T_e^{D(k)}$, and sample sizes, $n_e^{R(k)}$ and $n_e^{D(k)}$, for the experimental group by the randomized trial method and DS method.

3. Given the sample size $n_e^{R(k)}$ or $n_e^{D(k)}$, generate null experimental data from exponential distribution with hazard λ_c , alternative experimental data from exponential distribution with hazard λ_e and censoring times from $\text{uniform}(\tau, \tau + T_e^{(k)})$. Repeat this step L times to obtain L pairs of null and alternative experimental datasets.
4. For each pair of null and alternative experimental datasets, compute estimated hazard for the experimental group under the null and alternative hypotheses by $\hat{\lambda}_e^{0(k,l)} = d_e^{0(k,l)} / x_e^{0(k,l)}$ and $\hat{\lambda}_e^{1(k,l)} = d_e^{1(k,l)} / x_e^{1(k,l)}$, respectively, where the superscripts 0 and 1 indicate that the estimates are obtained under the null and alternative hypotheses. Compute test statistics

$$W^{0(k,l)} = \frac{\log \hat{\lambda}_c^{(k)} - \log \hat{\lambda}_e^{0(k,l)}}{\sqrt{1/d_e^{0(k,l)} + 1/d_c}}, \quad W^{1(k,l)} = \frac{\log \hat{\lambda}_c^{(k)} - \log \hat{\lambda}_e^{1(k,l)}}{\sqrt{1/d_e^{1(k,l)} + 1/d_c}}.$$

5. Evaluate the conditional type I error given $\hat{\lambda}_c^{(k)}$ by

$$\alpha^{(k)} = \sum_{l=1}^L I(W^{0(k,l)} > Z_{1-\alpha}) / L,$$

and the conditional power given $\hat{\lambda}_c^{(k)}$ by

$$pw^{(k)} = \sum_{l=1}^L I(W^{1(k,l)} > Z_{1-\alpha}) / L.$$

Let $K = L = 1000$. Under specification $\alpha = 0.05$, $1 - \beta = 0.8$, $\lambda_c = 0.0578$ corresponding to a median survival time of 12 months, $n_c = 100$, $d_c = 50$, $\delta = 1.5$, $r = 3$ patients per month, and $\tau = 12$ months, Figure 1 plots the histograms of the conditional type I error and the conditional power, based on all the simulated data and the randomized trial method. It shows that the distributions of the conditional type I error and the conditional power are highly skewed. Particularly, the conditional type I errors have a mean of 0.047 and a median of 0.001, and the conditional powers have a mean of 0.782 and a median of 0.943. For the DS method, the distributions of the conditional type I error and the conditional power are also highly skewed, as shown by the corresponding histograms in Figure 2. The conditional type I errors have a mean of 0.052 and a median of 0.013, and the conditional powers have a mean of 0.673 and a median of 0.756. Figure 3 plots the conditional power based on the DS method versus difference between the estimated hazard $\hat{\lambda}_c$ and the true λ_c in the HC group. It demonstrates that the conditional power roughly achieves the nominal level of 80% when the difference is 0. The conditional power is higher than 80% if the difference is less than 0 and lower than 80% if the difference is greater than 0. In Figure 4, we plot the required sample sizes for the experimental group versus estimated hazard $\hat{\lambda}_c$ with $n_c = 60$, 100 and 50% failure rate in the HC group, by the DS method (solid lines) and the randomized trial method

(dashed line). It shows that the sample size declines as the observed hazard increases, or the number of patients or failures in the HC group increases. Notice that the sample size by the randomized trial method is not available for $n_c = 60$ because $E(d_c)(\log \delta)^2 - (Z_{1-\alpha} + Z_{1-\beta})^2 \approx -1.218 < 0$ in (3), which suggests there is a limit to the extent to which accrual to the experimental group can compensate for lack of patients and failures in the HC group. Although not shown here, the DS method encounters the same problem because certain condition needs to be satisfied to ensure the existence of solution to equation (4) for determining $E(d_c)$.

3 Sample size controlling percentiles

Theoretical results of distributional properties of the conditional type I error and conditional power given the realization of the HC data, under the randomized trial method and DS method are described as follows, whose proof is provided in Appendix A.1.

Theorem 1

Let α be the nominal type I error and $1 - \beta$ be the nominal power. (a) The randomized trial method controls the mean of the conditional type I error at α and the mean of the conditional power at $1 - \beta$. (b) The DS method controls the mean of the conditional type I error at α and the median of the conditional power at $1 - \beta$.

For random variables with extremely skewed distributions, it is usually more desirable to make decisions based on a location parameter such as percentile rather than the mean. Controlling the conditional type I error and power by percentiles is appealing when their distributions are extremely skewed. The idea of controlling for percentiles of random power when there is uncertainty in some design parameters has been discussed in literature. Kieser and Wassmer (1996) suggested that the use of $100(1 - \gamma)\%$ upper one-sided confidence limit on the population variance would lead to a sample size which guarantees that the power will be greater than the nominal level with a probability of $1 - \gamma$. Glueck and Muller (2003) studied power analysis with a random baseline covariate in linear models, and introduced the concepts of conditional power as power calculated for a particular value of the random covariate, quantile power as the conditional power value that corresponds to the test statistics value of a specified probability, and unconditional power as the average power over all possible realizations of the random covariate. Since the distributions of the conditional type I error and power over all the random realizations of HC data are extremely skewed, we propose a flexible sample size formula to control arbitrary percentiles of $\alpha^{(k)}$ and $p_w^{(k)}$ for HC studies with survival outcomes.

Theorem 2

Suppose in an HC study, the goal is to control the $(1 - p_q)$ -th percentile of the conditional power at $1 - \beta$, and the p_h -th percentile of the conditional type I error at α . Then the required sample size is $n_c^ = rT_e^*$, where the accrual duration T_e^* is obtained by solving the following equation*

$$E^*(d_e) = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{\{\log \delta - (Z_{p_h} + Z_{p_q}) \sqrt{E(d_c)^{-1}}\}^2} = r[T_e^* - \exp(-\tau \lambda_e) \{1 - \exp(-T_e^* \lambda_e)\} / \lambda_e], \quad (5)$$

and the null hypothesis is rejected if

$$W^* = \frac{\log \hat{\lambda}_c - \log \hat{\lambda}_e - Z_{p_h} \sqrt{E(d_e)^{-1}}}{\sqrt{E^*(d_e)^{-1}}} > Z_{1-\alpha}.$$

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

$$\log \delta - (Z_{p_h} + Z_{p_q}) \sqrt{E(d_c)^{-1}} > 0 \text{ holds.}$$

The proof of Theorem 2 is presented in Appendix A.2. Given a particular HC dataset, the conditional type I error and power depend on the random difference between true hazard λ_c and estimated hazard $\hat{\lambda}_c$ in the HC group, as illustrated by simulations in Section 2.3.

Theorem 2 provides a formula to determine sample size n_e^* , under which $P(\alpha^* < \alpha) = p_h$ and $P(pw^* > 1 - \beta) = p_q$. This suggests that the conditional type I error would be smaller than the nominal level α with probability p_h and the conditional power would be greater than the nominal level $1 - \beta$ with probability p_q over all the possible HC data realizations. Thus, we can be confident that the operational characteristics (conditional type I error and power) of an HC study will be more desirable than the nominal levels with pre-specified probabilities p_h and p_q . The proposed method for determining sample size is flexible by allowing one to control arbitrary percentiles of the conditional type I error and power. By inspecting formula (5), one can see that the accrual requirement for the experimental group computed as n_e^* declines as (i) the accrual rate r decreases, (ii) the hazard ratio δ increases, (iii) the controlling percentiles (p_h, p_q) decrease, (iv) the number of failures in the HC group d_c increases, and (v) the hazard in the HC group λ_c increases. Further, as pointed out by Dixon and Simon (1988), there are situations where accrual to the experimental group can not compensate for lack of patients in the control group. For example, no number of new patients is large enough to simultaneously ensure that the conditional type I error would be smaller than 0.05 with probability 0.8 and the conditional power would be greater than 0.8 with probability 0.7 for detecting a hazard ratio $\delta = 1.5$, when only 10 failures are observed for the HC group. This is due to $\log \delta - (Z_{p_h} + Z_{p_q}) \sqrt{E(d_c)^{-1}} \approx -0.027 < 0$ in this case.

4 Simulation

Simulations are conducted to assess performance of the proposed method, considering different combinations of p_h and p_q that range from 0.5 to 0.9. In scenario 1, we assume an ideal case where true HC hazard rate λ_c and hazard ratio δ are known, and in scenario 2, we assume a more realistic case where only the hazard ratio δ is known. Table 1 summarizes simulation results under the ideal scenario and specification $\alpha = 0.05$, $1 - \beta = 0.8$, $\lambda_c = 0.0578$, $n_c = 100$, $d_c = 50$, $\delta = 1.5$, $r = 3$ and $\tau = 12$. It includes the estimated sample size n_e^* , the empirical means of the conditional type I error and power,

$\bar{\alpha}^* = \sum_{k=1}^K \alpha^{*(k)} / K$ and $\bar{pw}^* = \sum_{k=1}^K pw^{*(k)} / K$, and the empirical percentiles, $\hat{p}_h = \sum_{k=1}^K I(\alpha^{*(k)} < \alpha) / K$ and $\hat{p}_q = \sum_{k=1}^K I(pw^{*(k)} > 1 - \beta) / K$, where $\alpha^{*(k)}$ and $pw^{*(k)}$ are the conditional type I error and conditional power respectively, under samples size n_e^* for the k th replication, and $K = 1000$ is the total number of replications.

Table 2 presents simulation results under the realistic scenario, where the true hazard rate λ_c is unknown and replaced by an estimated value $\hat{\lambda}_c$ from a random realization of HC data, and thus the sample size n_e^* becomes random. The mean and standard deviation of the sample sizes under each combination of (p_h, p_q) are included in Table 2. The sample sizes are symmetric with respect to p_h and p_q in both tables, due to exchangeability of p_h and p_q in formula (5). As expected, the estimated sample size increases as the controlling percentiles p_h and p_q increase. Compared with Table 1, the sample sizes estimated based on the estimated $\hat{\lambda}_c$ in Table 2 center around those estimated based on the true hazard rate λ_c . Generally, the empirical means of the conditional type I error and power in Table 2 are in good agreement with those in Table 1, and the empirical percentiles, $(\hat{p}_h, \hat{p}_q)$, are close to their nominal levels, which suggests that the proposed approach performs reasonably well under the realistic scenario. Particularly, as shown in Table 1, under $(p_h, p_q) = (0.8, 0.7)$, we are confident that by enrolling 182 subjects, the conditional type I error would be smaller than 0.05 with probability 0.8, and the conditional power would be greater than 0.8 with probability 0.7.

Additionally, we examine how the required sample size for the experimental group varies under different configurations of accrual rate, difference to be detected and historical controls. When the accrual rate reduces to $r = 1$ patient per month, the estimated mean and standard deviation of sample size n_e^* are 154.01 and 3.19, respectively, under $(p_h, p_q) = (0.8, 0.7)$ and the realistic scenario with other design parameters unchanged. When the hazard ratio increases to $\delta = 1.75$, the corresponding estimated mean and standard deviation of sample size are 82.79 and 5.55, respectively. When the number of patients and failures in the HC group decrease to $n_c = 60$ and $d_c = 30$, the estimated mean and standard deviation of sample size are 302.31 and 12.12, respectively. Lastly, when the hazard in the HC group increases to $\lambda_c = 0.1155$, corresponding to a median survival time of 6 months, the estimated mean and standard deviation of sample size are 153.02 and 4.03, respectively.

5 Application

An investigator wishes to design an advanced non-small cell lung cancer (NSCLC) clinical trial to compare the progression-free survival of patients under a new chemotherapy regimen (E) with those under a control treatment (C) of erlotinib. A previous study shows that the median progression-free survival was 2.6 months, corresponding to a hazard rate of $\lambda_c = 0.267$, in $n_c = 68$ patients who received a control treatment (erlotinib) and $d_c = 50$ patients had disease progression (Spigel et al., 2013). Data will be collected for a new chemotherapy treatment regimen through a prospective clinical trial. Since the observed HC group is a realization from the underlying HC distribution, we need to take the uncertainty about the HC data into consideration in the HCT design, and control the conditional power and type I

error at desired levels. The investigator wants to estimate the sample size needed to have 80% power to detect a ratio of hazard rates of $\delta = 1.5$, using a one-sided, 5% significance level test with a follow-up period of $\tau = 12$ months. Recent experience indicates that new eligible patients will be accrued at the rate of $r = 5$ patients per month.

As discussed previously, the conditional power and type I error over all the possible realizations of HC data have highly skewed distributions, so it is more reasonable to control their percentiles instead of means. Therefore, we estimate the number of patients needed to detect a hazard ratio of 1.5 comparing a control treatment to a new chemotherapy regimen, controlling the 70th percentile ($p_h = 0.7$) of the conditional type I error at 5%, and the 30th percentile ($p_q = 0.7$) of the conditional power at 80%. The number of patients required in the new chemotherapy regimen group is $n_e^* = 97$, at which the conditional power of the clinical trial will be greater than 80% with probability of 0.7, and the conditional type I error will be smaller than 5% with probability of 0.7. Additionally, under the randomized trial method that controls the mean of the conditional power at 80% and the mean of the conditional type I error at 5%, the required sample size is $n_e^R = 156$. The sample size estimated under the DS method is $n_e^D = 73$, which controls the median of the conditional power at 80% and the mean of the conditional type I error at 5%.

6 Discussion

This paper, to our knowledge, is the first attempt to systematically investigate the performances of standard methods for HC design with survival outcomes. It proposes a more flexible alternative method of determining the required sample size, controlling for arbitrary percentiles of the conditional power and type I error, and accommodates the extreme skewness in the distributions of the conditional power and type I error. The controlling percentiles can be set according to the requirement of specific trials. It is noted that in the NSCLC clinical trial example, the sample size by the randomized trial method, $n_e^R = 156$, is approximately equal to n_e^* calculated under $p_h = 0.73$, $p_q = 0.73$. If the investigator considers that controlling the means of the conditional power and type I error at desired levels is too stringent for random variables with skewed distribution, it is reasonable to set lower controlling percentiles ($p_h = 0.7$, $p_q = 0.7$).

We focus on the case with no additional follow-up of remaining controls for convenience of discussion. This situation arises frequently when the investigator conducting the current study had not collected the control group data, or it is not possible for further follow-up of control group subjects. On the other hand, there may be situations when one can obtain further follow-up on the control group subjects who have not yet failed, and still treat all newly recruited subjects with the experimental treatment. In this case, the expected number of failures in control group becomes $E(d_c)' = d_c + y_c[1 - \exp\{-(T + \tau)\lambda_c\}]$, where d_c is the number of already observed failures and y_c is the number of patients still at risk in the HC group. The asymptotic variance of $\log(\hat{\lambda}_c/\hat{\lambda}_e)$ becomes

$$\frac{y_c[1 - \exp\{-(T + \tau)\lambda_c\}]}{(d_c + y_c[1 - \exp\{-(T + \tau)\lambda_c\}])^2} + \frac{1}{E(d_e)}.$$

Accordingly, the sample size formulas by various methods need to be adjusted. For HC studies with survival outcomes, the required accrual to the experimental group and censoring mechanism are closely related. Under the exponential survival and uniform censoring distribution assumptions, the proportion of censoring can be determined by $P(\text{censoring}) = \exp(-\tau\lambda_e)\{1 - \exp(-T\lambda_e)\}/T\lambda_e$. Thus, as the accrual period, T , increases, the proportion of censoring decreases. Our simulation experience shows that the average censoring rate is around 30% for $n_c = 100$ and $d_c = 50$, and 24% for $n_c = 60$ and $d_c = 30$ under the DS method, which is consistent with the theoretical result.

The proposed method is most useful for a moderate sample size in the HC group. In fact, when the number of patients in the HC group, n_c , is very large, a comparison with the hazard rate of the HC group is similar to comparing with a fixed constant, and thus, the sample size calculation is not as critical because all methods lead to similar answer. On the other hand, when n_c is small, accrual to the experimental groups may not be able to compensate lack of patients and failures in the HC group, as shown in the previous sections. HCTs are designed under the assumption that the historical controls are from a similar population with the patients in the experimental group. In reality, patient selection factors used for the historical control group may be quite different from those used for the current treatment group. Changes over time in diagnosis and evaluation criteria can also bias the results of comparisons. Therefore, HCTs may not produce definitive results as RCTs do (Byar et al., 1976). Pocock (1976) provided the requirements for a valid HCT, and Gehan (1986) discussed establishing comparability between randomized concurrent control groups and historical control groups. Nevertheless, there are situations where a RCT is not feasible or necessary. For example, it may be difficult to convince clinicians to enter their patients into a randomized trial when the standard treatment is known to have no effect. An HCT may be considered acceptable and the sample size method proposed in this paper could be used to properly design the HCT to give sufficient power against reasonable alternatives.

Acknowledgement

This work was supported in part by the Cancer Center Support Grant (5P30CA142543) from the National Cancer Institute awarded to the Harold C. Simmons Cancer Center at the University of Texas Southwestern Medical Center. The authors thank the editor, the associate editor and two reviewers for their constructive comments that have greatly improved the initial version of this paper.

Appendix

A.1 Proof of Theorem 1

Proof

By the large-sample normal theory, we have $\log \hat{\lambda}_c \sim N(\log \lambda_c, 1/E(d_c))$ for the HC group. For the randomized trial method, we use

$\log \hat{\lambda}_e^{0R} \sim N(\log \lambda_c, 1/E(d_e^R))$ and $\log \hat{\lambda}_e^{1R} \sim N(\log \lambda_e, 1/E(d_e^R))$ to denote statistics from the experimental group under the null and alternative hypotheses, respectively. Similarly, we

define $\log \hat{\lambda}_e^{0D} \sim N(\log \lambda_c, 1/E(d_e^D))$ and $\log \hat{\lambda}_e^{1D} \sim N(\log \lambda_e, 1/E(d_e^D))$ for the DS method. Let

$f(\cdot)$ denote the density function of a random variable. Under the randomized trial method, the mean of the conditional type I error over all iterations ($k = 1, \dots, K$) is

$$\begin{aligned}\text{Mean}(\alpha_R^{(k)}) &= \int I\left(\frac{\log \hat{\lambda}_c - \log \hat{\lambda}_e^{0R}}{\sqrt{1/E(d_e^R) + 1/E(d_c)}} > Z_{1-\alpha}\right) f(\hat{\lambda}_e^{0R}) f(\hat{\lambda}_c) d\hat{\lambda}_e^{0R} d\hat{\lambda}_c \\ &= \int I(W^{0R} > Z_{1-\alpha}) f(W^{0R}) dW^{0R} = P(W^{0R} > Z_{1-\alpha}) = \alpha,\end{aligned}$$

where $W^{0R} = (\log \hat{\lambda}_c - \log \hat{\lambda}_e^{0R}) / \sqrt{1/E(d_e^R) + 1/E(d_c)}$ follows $N(0, 1)$ distribution.

Similarly, it can be shown that $\text{Mean}(\alpha_D^{(k)}) = \alpha$ under the DS method.

Next, we show that the randomized trial method controls the mean of the conditional power at $1 - \beta$,

$$\begin{aligned}\text{Mean}(pw_R^{(k)}) &= \int I\left(\frac{\log \hat{\lambda}_c - \log \hat{\lambda}_e^{1R}}{\sqrt{1/E(d_e^R) + 1/E(d_c)}} > Z_{1-\alpha}\right) f(\hat{\lambda}_e^{1R}) f(\hat{\lambda}_c) d\hat{\lambda}_e^{1R} d\hat{\lambda}_c \\ &= \int I(W^{1R} > Z_{1-\alpha} - \frac{\log \delta}{\sqrt{1/E(d_e^R) + 1/E(d_c)}}) f(W^{1R}) dW^{1R} \\ &= \int I(W^{1R} > -Z_{1-\beta}) f(W^{1R}) dW^{1R} = 1 - \beta,\end{aligned}$$

where $W^{1R} = (\log \hat{\lambda}_c - \log \hat{\lambda}_e^{1R} - \log \delta) / \sqrt{1/E(d_e^R) + 1/E(d_c)}$ follows $N(0, 1)$ distribution.

The third equality is obtained by plugging in formula (3) for $E(d_e^R)$.

Lastly, we prove that the DS method controls the median of the conditional power at $1 - \beta$. We have

$$\begin{aligned}pw_D^{(k)} &= P\left(\frac{\log \hat{\lambda}_c - \log \hat{\lambda}_e^{1D}}{\sqrt{1/E(d_e^D) + 1/E(d_c)}} > Z_{1-\alpha} \mid \hat{\lambda}_c = \hat{\lambda}_c^{(k)}\right) \\ &= P\left(\frac{(\log \hat{\lambda}_c - \log \hat{\lambda}_e^{1D}) - (\log \hat{\lambda}_c - \log \lambda_e)}{\sqrt{1/E(d_e^D)}} > \frac{Z_{1-\alpha} \sqrt{1/E(d_e^D) + 1/E(d_c)}}{\sqrt{1/E(d_e^D)}} - \frac{\log \hat{\lambda}_c - \log \lambda_e}{\sqrt{1/E(d_e^D)}} \mid \hat{\lambda}_c = \hat{\lambda}_c^{(k)}\right) \\ &= 1 - \phi\left(\frac{Z_{1-\alpha} \sqrt{1/E(d_e^D) + 1/E(d_c)}}{\sqrt{1/E(d_e^D)}} - \frac{\log \hat{\lambda}_c - \log \lambda_e}{\sqrt{1/E(d_e^D)}} \mid \hat{\lambda}_c = \hat{\lambda}_c^{(k)}\right).\end{aligned}$$

Obviously, $pw_D^{(k)}$ is a monotone increasing function in $\hat{\lambda}_c^{(k)}$. The expected number of failures $E(d_e^D)$ by the DS method is obtained as the solution to $pw_D^{(k)} = 1 - \beta$ when $\hat{\lambda}_c^{(k)} = \lambda_c$, the median of its distribution. Since $pw_D^{(k)}$ is a monotone function of $\hat{\lambda}_c^{(k)}$ and equals $1 - \beta$ at the median of $\hat{\lambda}_c^{(k)}$, we have $\text{Median}(pw_D^{(k)}) = 1 - \beta$.

A.2 Proof of Theorem 2

Proof

We first demonstrate that based on W^* , the p_h -th percentile of the conditional type I error is controlled at α . For a given $\log \hat{\lambda}_c$, the conditional type I error is

$$\alpha^* = P\left(\frac{\log \hat{\lambda}_c - \log \hat{\lambda}_e^0 - Z_{p_h} \sqrt{E(d_c)^{-1}}}{\sqrt{E^*(d_e)^{-1}}} > Z_{1-\alpha}\right) = 1 - \phi\left(Z_{1-\alpha} - \frac{\log \hat{\lambda}_c - (\log \lambda_c + Z_{p_h} \sqrt{E(d_c)^{-1}})}{\sqrt{E^*(d_e)^{-1}}}\right).$$

When $\hat{\lambda}_c = \log \lambda_c + Z_{p_h} \sqrt{E(d_c)^{-1}}$, which is the p_h -th percentile of $\log \hat{\lambda}_c \sim N(\log \lambda_c, E(d_c)^{-1})$, the conditional type I error $\alpha^* = \alpha$. Since α^* is a monotone increasing function in $\log \hat{\lambda}_c$, we have $P(\alpha^* < \alpha) = P(\log \hat{\lambda}_c < \log \lambda_c + Z_{p_h} \sqrt{E(d_c)^{-1}}) = p_h$.

Next, we find the sample size $n_e^* = rT_e^*$ and corresponding accrual duration T_e^* which control the $(1 - p_q)$ -th percentile of the conditional power at $1 - \beta$. The conditional power given $\log \hat{\lambda}_c$ is

$$pw^* = P\left(\frac{\log \hat{\lambda}_c - \log \hat{\lambda}_e^1 - Z_{p_h} \sqrt{E(d_c)^{-1}}}{\sqrt{E^*(d_e)^{-1}}} > Z_{1-\alpha}\right) = 1 - \phi\left(Z_{1-\alpha} - \frac{\log \hat{\lambda}_c - \log \lambda_c + \log \delta - Z_{p_h} \sqrt{E(d_c)^{-1}}}{\sqrt{E^*(d_e)^{-1}}}\right), \quad (6)$$

and pw^* is monotonically increasing with $\log \hat{\lambda}_c$. Based on the monotonicity, if we set $pw^* = 1 - \beta$ at $\log \hat{\lambda}_c = \log \lambda_c - Z_{p_q} \sqrt{E(d_c)^{-1}}$, which is the $(1 - p_q)$ -th percentile of $\log \hat{\lambda}_c$, the $(1 - p_q)$ -th percentile of the conditional power pw^* will be controlled at $1 - \beta$, because

$$P(pw^* > 1 - \beta) = P(\log \hat{\lambda}_c > \log \lambda_c - Z_{p_q} \sqrt{E(d_c)^{-1}}) = p_q.$$

Lastly, let $pw^* = 1 - \beta$ and $\log \hat{\lambda}_c = \log \lambda_c - Z_{p_q} \sqrt{E(d_c)^{-1}}$ in (6), we have

$$1 - \beta = 1 - \phi\left(Z_{1-\alpha} - \frac{\log \delta - (Z_{p_h} + Z_{p_q}) \sqrt{E(d_c)^{-1}}}{\sqrt{E^*(d_e)^{-1}}}\right),$$

from which $E^*(d_e)$ can be determined. Combining with

$E^*(d_e) = r[T_e^* - \exp(-\tau \lambda_e)\{1 - \exp(-T_e^* \lambda_e)\}/\lambda_e]$, we can compute the accrual duration T_e^* and the sample size required for the experimental group as $n_e^* = rT_e^*$. The condition

$\log \delta - (Z_{p_h} + Z_{p_q}) \sqrt{E(d_c)^{-1}} > 0$ needs to be satisfied due to the positive constraint on $\sqrt{E^*(d_e)}$ in (5).

References

1. Byar DP, Simon RM, Friedewald WT, Schlesselman JJ, DeMets DL, Ellenberg JH, Gail MH, Ware JH. Randomized clinical trials: Perspectives on some new ideas. *New England Journal of Medicine*. 1976; 295:74–80. [PubMed: 775331]
2. Buchs AE, Silverman BG. Incidence of malignancies in patients with diabetes mellitus and correlation with treatment modalities in a large Israeli health maintenance organization: a historical cohort study. *Metabolism*. 2011; 60:1379–1385. [PubMed: 21696791]
3. Chen M, Tse LA, Au KF, Yu TS, Wang X, Lao X, Au JS. Mesothelioma and lung cancer mortality: A historical cohort study among asbestosis workers in Hong Kong. *Lung Cancer*. 2012; 76:165–170. [PubMed: 22129857]
4. Dixon DO, Simon RM. Sample size considerations for studies comparing survival curves using historical controls. *Journal of Clinical Epidemiology*. 1988; 41:1209–1213. [PubMed: 3062140]
5. Emrich LJ. Required duration and power determinations for historically controlled studies of survival times. *Statistics in Medicine*. 1989; 8:153–160. [PubMed: 2704897]
6. Gehan EA. The evaluation of therapies: Historical control studies. *Statistics in Medicine*. 1984; 3:315–324. [PubMed: 6528131]
7. Gehan EA. Randomized or historical control groups in cancer clinical trials: Are historical controls valid? *Journal of Clinical Oncology*. 1986; 4:1024–1025. [PubMed: 3723163]
8. Glueck DH, Muller KE. Adjusting power for a baseline covariate in linear models. *Statistics in Medicine*. 2003; 22:2535–2551. [PubMed: 12898543]
9. Ibrahim JG, Ryan LM, Chen M-H. Using historical controls to adjust for covariates in trend tests for binary data. *Journal of the American Statistical Association*. 1998; 93:1282–1293.
10. Keiding N. Historical controls and modern survival analysis. *Lifetime Data Analysis*. 1995; 1:1–25.
11. Kepner J, Wackerly D. Some observations on the makuch/simon approach to sample size determination in clinical trials with historical controls. *Communications in Statistics: Simulation and Computation*. 2001; 30:611–621.
12. Kieser M, Wassmer G. On the use of the upper confidence limit for the variance from a pilot sample for sample size determination. *Biometrical Journal*. 1996; 38:941–949.
13. Korn EL, Freidlin B. Conditional power calculations for clinical trials with historical controls. *Statistics in Medicine*. 2006; 25:2922–2931. [PubMed: 16479548]
14. Lee JJ, Tseng C. Uniform power method for sample size calculation in historical control studies with binary response. *Controlled Clinical Trials*. 2001; 22:390–400. [PubMed: 11514040]
15. 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; 101:148–155. [PubMed: 18344473]
16. Makuch RW, Simon RM. Sample size considerations for non-randomized comparative studies. *Journal of Chronic Diseases*. 1980; 33:175–181. [PubMed: 7354106]
17. Pinsky CM. Experience with historical control studies in cancer immunotherapy. *Statistics in Medicine*. 1984; 3:325–329. [PubMed: 6528132]
18. Pocock SJ. The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*. 1976; 29:175–188. [PubMed: 770493]
19. Spigel DR, Ervin TJ, Ramlau RA, Daniel DB, Goldschmidt JH Jr, Blumenschein GR Jr, Krzakowski MJ, Robinet G, Godbert B, Barlesi F, Govindan R, Patel T, Orlov SV, Wertheim MS, Yu W, Zha J, Yauch RL, Patel PH, Phan SC, Peterson AC. Randomized phase II trial of Onartuzumab in combination with erlotinib in patients with advanced non-small-cell lung cancer. *Journal of Clinical Oncology*. 2013; 31:4105–4114. [PubMed: 24101053]
20. Zhang S, Cao J, Ahn C. Calculating sample size in trials using historical controls. *Clinical Trials*. 2010; 7:343–353. [PubMed: 20573638]
21. Zhang S, Cao J, Ahn C. Sample size calculation for studies comparing binary outcomes using historical controls. *Biometrical Journal*. 2013; 55:190–202. [PubMed: 23335222]

Figure 1.
The distributions of conditional type I error and power, based on all simulated data and randomized trial method.

Figure 2.
The distributions of conditional type I error and power, based on all simulated data and DS method.

Figure 3.

The conditional power based on DS method versus difference between estimated and true hazards in the HC group.

Figure 4.

Sample size for the experimental group versus estimated hazard in the HC group with $n_c = 60, 100$ and 50% failure rate, by DS method (solid lines) and randomized trial method (dashed line).

Table 1

Simulation results under the ideal scenario and sample size controlling percentiles (p_h, p_q).

		p_q				
		0.5	0.6	0.7	0.8	0.9
0.5	n_e^*	66	76	90	113	167
	$(\bar{\alpha}^*, \overline{pw}^*)$	(0.10,0.74)	(0.12,0.79)	(0.13,0.85)	(0.16,0.90)	(0.21,0.94)
	$[p_h^*, p_q^*]$	[0.54,0.50]	[0.56,0.59]	[0.53,0.74]	[0.53,0.81]	[0.50,0.92]
0.6	n_e^*	76	89	108	140	221
	$(\bar{\alpha}^*, \overline{pw}^*)$	(0.09,0.75)	(0.10,0.79)	(0.10,0.84)	(0.13,0.90)	(0.17,0.95)
	$[p_h^*, p_q^*]$	[0.60,0.53]	[0.63,0.63]	[0.61,0.72]	[0.60,0.83]	[0.59,0.92]
0.7	n_e^*	90	108	134	182	323
	$(\bar{\alpha}^*, \overline{pw}^*)$	(0.06,0.74)	(0.08,0.78)	(0.10,0.84)	(0.11,0.89)	(0.13,0.94)
	$[p_h^*, p_q^*]$	[0.71,0.49]	[0.71,0.61]	[0.67,0.74]	[0.68,0.82]	[0.71,0.91]
0.8	n_e^*	113	140	182	269	608
	$(\bar{\alpha}^*, \overline{pw}^*)$	(0.04,0.71)	(0.05,0.77)	(0.06,0.82)	(0.09,0.88)	(0.11,0.94)
	$[p_h^*, p_q^*]$	[0.83,0.50]	[0.81,0.62]	[0.81,0.71]	[0.79,0.81]	[0.80,0.91]
0.9	n_e^*	167	221	323	608	3395
	$(\bar{\alpha}^*, \overline{pw}^*)$	(0.03,0.69)	(0.03,0.75)	(0.04,0.83)	(0.05,0.86)	(0.09,0.92)
	$[p_h^*, p_q^*]$	[0.91,0.51]	[0.91,0.61]	[0.89,0.74]	[0.89,0.80]	[0.89,0.91]

The estimated sample size n_e^* , the empirical means of the conditional type I error and power, $(\bar{\alpha}^*, \overline{pw}^*)$ and the empirical percentiles, $[p_h^*, p_q^*]$. Assume that true hazard rate λ_c and hazard ratio δ are known.

Table 2

Simulation results under the realistic scenario and sample size controlling percentiles (p_h, p_q).

		p_q				
		0.5	0.6	0.7	0.8	0.9
0.5	$n_e^*(sd)$	66.04 (4.28)	76.15 (4.90)	90.08 (5.51)	112.87 (6.57)	166.39 (7.62)
	$(\bar{\alpha}^*, \overline{pw}^*)$	(0.10,0.75)	(0.12,0.80)	(0.14,0.85)	(0.17,0.90)	(0.20,0.95)
	$[p_h^*, p_q^*]$	[0.53,0.52]	[0.54,0.63]	[0.52,0.74]	[0.51,0.84]	[0.49,0.92]
0.6	$n_e^*(sd)$	76.52 (4.91)	89.13 (5.39)	107.95 (6.18)	139.56 (6.91)	220.69 (8.57)
	$(\bar{\alpha}^*, \overline{pw}^*)$	(0.08,0.72)	(0.10,0.80)	(0.11,0.84)	(0.13,0.90)	(0.17,0.95)
	$[p_h^*, p_q^*]$	[0.64,0.48]	[0.61,0.63]	[0.60,0.72]	[0.62,0.83]	[0.61,0.92]
0.7	$n_e^*(sd)$	90.26 (5.49)	107.99 (6.06)	134.48 (6.84)	181.68 (8.15)	323.32 (9.52)
	$(\bar{\alpha}^*, \overline{pw}^*)$	(0.07,0.73)	(0.07,0.78)	(0.08,0.83)	(0.11,0.89)	(0.14,0.95)
	$[p_h^*, p_q^*]$	[0.71,0.53]	[0.73,0.61]	[0.73,0.69]	[0.68,0.81]	[0.71,0.93]
0.8	$n_e^*(sd)$	113.39 (6.15)	139.70 (6.88)	182.24 (7.93)	268.84 (9.13)	607.95 (10.05)
	$(\bar{\alpha}^*, \overline{pw}^*)$	(0.04,0.71)	(0.05,0.77)	(0.06,0.81)	(0.07,0.88)	(0.11,0.94)
	$[p_h^*, p_q^*]$	[0.83,0.50]	[0.82,0.62]	[0.82,0.70]	[0.80,0.82]	[0.79,0.91]
0.9	$n_e^*(sd)$	166.84 (7.63)	220.95 (8.36)	322.70 (9.51)	608.08 (9.77)	3394.97 (9.99)
	$(\bar{\alpha}^*, \overline{pw}^*)$	(0.03,0.68)	(0.03,0.75)	(0.05,0.82)	(0.05,0.87)	(0.09,0.94)
	$[p_h^*, p_q^*]$	[0.90,0.50]	[0.91,0.61]	[0.88,0.74]	[0.89,0.82]	[0.87,0.92]

The estimated mean (standard deviation, sd) of sample size n_e^* , the empirical means of the conditional type I error and power, $(\bar{\alpha}^*, \overline{pw}^*)$ and the empirical percentiles, $[p_h^*, p_q^*]$. Assume that true hazard rate λ_C is unknown and hazard ratio δ is known.