Published in final edited form as:

Pharm Stat. 2016 September ; 15(5): 405–411. doi:10.1002/pst.1756.

Survival Trial Design and Monitoring Using Historical Controls

Jianrong Wu and Xiaoping Xiong

Department of Biostatistics, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, USA

Abstract

In this paper, we propose a multistage group sequential procedure to design survival trials using historical controls. The formula for the number of events required for historical control trial designs is derived. Furthermore, a transformed information time is proposed for trial monitoring. An example is given to illustrate the application of the proposed methods to survival trial designs using historical controls.

Keywords

group sequential trial design; historical control; log-rank test; information time; sample size; time-to-event

1 Introduction

Randomized clinical trials (RCTs) are the gold standard for clinical trials comparing a treatment group to a control group. When randomization is not feasible because of ethical concerns, patient preference, or regulatory acceptability, historical control trials (HCTs) are the alternative. A pronounced argument for using HCTs is that all patients can receive the new treatment. The major benefits of the HCTs include their contribution to medical knowledge and potential cost savings in sample size and length of study. However, HCTs may suffer from patient selection bias because patients with more favorable prognoses may be more likely to be selected to receive the new treatment. The HCTs may also suffer from outcome evaluation bias due to changes in the experimental environment as well as technology, and have been criticized accordingly [1–6].

Despite of these facts, the HCTs has been widely applied in medical clinical research [7–13]. In randomized trials, factors such as age, race, and gender, that may be confounded with treatment effects are balanced with respect to their marginal or joint frequencies in the control and experimental groups through randomization. Obviously, an HCT lacks this mechanism and hence, in the application of HCTs, it is important to make sure that patient's eligibility criteria of the experimental group should be the same as or at least similar to the historical control study. If it is not possible to guarantee a minimum selection bias in the design phase, or it is difficult to implement the balance in the trial process, the potential confounding effects of factors with unbalanced frequencies should be investigated in later analyses.

Sample size calculations to design HCTs have been discussed by Makuch and Simon [14] for binary endpoints and by Dixon and Simon [15] and Emrich [16] for exponential survival endpoints. When assessing the power in these designs, these authors assumed the true control treatment effect to be equal to the observed effect in the historical control group. As Korn and Freidlin [17] pointed out that these popular methods do not preserve the power and type I error when considering the uncertainty in the historical control data. Recently, several studies have discussed sample size calculations for HCTs by taking into account the uncertainty in the true historical control treatment effect [18–20].

HCTs are often monitored by pre-planned interim analyses to stop accrual if patients in the experimental group have poorer outcomes than those in the historical control group. Monitoring clinical trials with historical controls poses a statistical problem of comparing two outcomes in a situation wherein data from the experimental group are sequentially collected and compared with all data from historical controls at each interim analysis. The widely used information time for survival trials, defined as the ratio of the number of events to be observed at an interim analysis to that the total number in the entire trial [21], is actually not appropriate for HCTs, especially when the sample size of historical data is substantially different from that of current data which is a case often seen. Few publications have discussed the monitoring of clinical trials using historical controls. For example, Chang et al. [18] proposed a two-stage design for binary outcomes and Xiong et al. [19] developed a multistage group sequential procedure for monitoring HCTs with binary, continuous, and survival endpoints. However, the information time calculation for survival endpoints applied in Xiong et al. [19] was based on the traditional ratio of event numbers mentioned above and are not accurate for such an application.

In this paper, we propose a multistage group sequential procedure to design survival trials that use historical controls. A sequential log-rank test using historical controls is proposed. Based on this sequential log-rank test, we derive sample size calculations for sequential HCT designs. Furthermore, a transformed information time is derived to define the sequential boundary based on the sequential conditional probability ratio test (SCPRT) proposed by Xiong [22]. The rest of the paper is organized as follows. In Section 2, a sequential log-rank test using historical controls is proposed. In Section 3, formulae for the number of events and sample size required for sequential trial designs using historical controls are derived. A transformed information time is proposed in Section 4. In Section 5, a multistage group sequential procedure is discussed. In Section 6, an example is given to illustrate the proposed methods. The discussion and concluding remarks are presented in Section 7.

2 Sequential Log-rank Test

Suppose there are two groups, a historical control group, and an experimental group, which are designated by group 1 and group 2, respectively. Assuming the survival distributions of the two groups satisfy the proportional hazards model

$$S_1(t)=[S_2(t)]^\delta, \quad (1)$$

where δ is the unknown hazard ratio of the historical control group versus experimental group. The hypothesis of improvement in the survival distribution of the treatment group compared to that of the historical control group can be expressed as

$$H_0: \delta \leq 1 \text{ vs. } H_1: \delta > 1. \quad (2)$$

A well-known test statistic for testing the above hypothesis is the log-rank test. We introduce a sequential log-rank test statistic for HCTs as follows. Suppose there are r distinct event (or failure) times, $t_{(1)} < t_{(2)} \cdots < t_{(r)}$, among the two groups. An event time is measured from the time of enrolling in the study to the time of an event for a patient. At an event time $t_{(j)}$, d_{1j} events occur for the historical control group and d_{2j} events occur for the experimental group, respectively, with n_{1j} and n_{2j} subjects being at risk in the two groups just prior to $t_{(j)}$. Thus, there are $d_j = d_{1j} + d_{2j}$ events at $t_{(j)}$ among a total of $n_j = n_{1j} + n_{2j}$ subjects. Let τ be the study duration of the experimental group and $t (< \tau)$ be the calendar time from the start of study enrollment to the interim look, and let $n_{2j}(t)$ and $d_{2j}(t)$ represent the part of n_{2j} and d_{2j} with calendar times of entry no later than t , respectively. Let $n_j(t) = n_{1j} + n_{2j}(t)$, $d_j(t) = d_{1j} + d_{2j}(t)$, and $e_{2j}(t) = n_{2j}(t)d_{2j}(t)/n_j(t)$. Then the log-rank score at an interim look at calendar time t is

$$U(t) = \sum_{j=1}^r \{d_{2j}(t) - e_{2j}(t)\}, \quad (3)$$

which is approximately normal with a mean of $\log(\delta) V(t)$ and variance of $V(t)$, where the variance $V(t)$ is given by

$$V(t) = \sum_{j=1}^r \frac{n_{1j}n_{2j}(t) d_j(t) (n_j(t) - d_j(t))}{n_j^2(t) (n_j(t) - 1)}.$$

Thus, the sequential log-rank test at calendar time t is given by

$$L(t) = U(t)/V^{1/2}(t), \quad (4)$$

which is approximately normal distributed with a mean of $\log(\delta) V^{1/2}(t)$ and variance of 1 under local alternative. Therefore, $B_{t^*} = U(t)/V^{1/2}(\tau) \sim N(\theta t^*, t^*)$ is approximately a Brownian motion with drift parameter $\theta = \log(\delta) V^{1/2}(\tau)$ [23–25], and information time $t^* = V(t)/V(\tau)$, where $V(\tau)$ is the value of $V(t)$ at $t = \tau$. The difference between a two-sample sequential log-rank test and a historical control sequential log-rank test is that at an interim analysis, the two-sample sequential log-rank test uses the data up to the calendar time t of both groups, whereas the sequential log-rank test for the HCTs uses all data from the historical control group but only data up to time t for the experimental group.

3 Sample Size Calculation

A key step in designing a survival clinical trial is to calculate the required number of events or sample size based on testing the hypothesis (2) under the proportional hazards model (1), which is equivalent to $\lambda_1(t) = e^\gamma \lambda_2(t)$, where $\lambda_1(t)$ and $\lambda_2(t)$ are the hazard functions of groups 1 and 2, respectively, and $\gamma = \log(\delta)$ is the log hazard ratio. Let ω_1 and $\omega_2 = 1 - \omega_1$ be the proportions of subjects assigned to treatment groups 1 and 2, respectively. Define $p_i(\tau)$ as the probability of a patient in the i^{th} group having an event during the study period $[0, \tau]$, and let $P(\tau) = \omega_1 p_1(\tau) + \omega_2 p_2(\tau)$ be the probability of failure for a patient randomly chosen from the two groups with probabilities of ω_1 and ω_2 . To derive the power and sample size calculation, we consider a sequence of local alternatives H_{1n} : $\gamma = \gamma_{1n} = b/\sqrt{n}$, where $b < \infty$. Under local alternative H_{1n} , Schoenfeld [26] has shown that the log-rank test $L(\tau)$ is asymptotically normal with unit variance and mean $b\sigma(\tau)$, where $\sigma^2(\tau) = \omega_1 \omega_2 P(\tau)$ is the asymptotic variance of the log-rank score $U(\tau)$. As $p_1(\tau) \simeq p_2(\tau)$ under H_{1n} with large n , we have

$$\left(\frac{1}{n_1 p_1(\tau)} + \frac{1}{n_2 p_2(\tau)} \right)^{-1} \simeq \omega_1 \omega_2 \{n_1 p_1(\tau) + n_2 p_2(\tau)\} = n \sigma^2(\tau).$$

Replacing the approximating variance $n\sigma^2(\tau)$ of $L(\tau)$ with $\sigma^2(\tau) = (D_1(\tau)^{-1} + D_2(\tau)^{-1})^{-1}$ where $D_i(\tau) = n_i p_i(\tau)$, and given a type I error of α , the power $(1 - \beta)$ of the log-rank test L satisfies the following:

$$1 - \beta = P(L(\tau) > z_{1-\alpha} | H_{1n}) \simeq \Phi(|\log(\delta)| \tilde{\sigma}(\tau) - z_{1-\alpha}),$$

Hence, the expected number of events in each group satisfies the following equation:

$$D_1(\tau)^{-1} + D_2(\tau)^{-1} = \frac{[\log(\delta)]^2}{(z_{1-\alpha/2} + z_{1-\beta})^2}. \quad (5)$$

This formula was first derived by Rubinstein et al. [28] based on the maximum likelihood test under the exponential model. Our simulations indicated that this formula works well for unbalanced trial designs under the fixed alternative H_1 .

Because the historical control data are obtained from previous trials, the total number of events $D_1 = D_1(\tau)$ for the historical control group is known. Therefore, to design a historical control trial, we need only calculate the number of events or the sample size for the experimental group by assuming no further follow-up for the historical control group or that the historical control data are frozen at the design stage. Thus, for HCTs, the sample size formula (5) can be written as

$$D_2^{-1}(\tau) + D_1^{-1} = \frac{[\log(\tau)]^2}{(z_{1-\alpha} + z_{1-\beta})^2}, \quad (6)$$

where D_1 is the total number of events observed in the historical control data, which is known, and $D_2(\tau)$ is the expected total number of events for the experimental group. The expected total number of events required for the experimental group based on the log-rank test $L(\tau)$ can be obtained by solving equation (6) as

$$D_2(\tau) = \left\{ \frac{[\log(\delta)]^2}{(z_{1-\alpha} + z_{1-\beta})^2} - D_1^{-1} \right\}^{-1}, \quad (7)$$

and the sample size for the experimental group is given by

$$N_2 = D_2(\tau) / p_2(\tau), \quad (8)$$

where $p_2(\tau)$ is the probability of a subject from the experimental group having an event during the study.

To calculate the sample size using formula (8), we have to calculate $p_2(\tau)$. We assume that subjects are recruited with a uniform distribution over the accrual period t_a and are followed for time t_f and that the study duration is $\tau = t_a + t_f$. We further assume that no subjects are lost to follow-up. Then the distribution of the censoring time is uniform over the interval $[t_f, t_a + t_f]$. Thus, $p_2(\tau)$ can be calculated by

$$p_2(\tau) = 1 - \frac{1}{t_a} \int_{t_f}^{t_a + t_f} S_2(t) dt, \quad (9)$$

where $S_2(t) = [S_1(t)]^{1/6}$, and $S_1(t)$ is the survival distribution of the historical control group, which can be estimated from the historical data by a parametric survival distribution, a Kaplan-Meier curve, or a spline version of the survival distribution without a specific model assumption [29]. This integration can also be approximated numerically using Simpson's rule,

$$p_2(\tau) = 1 - \frac{1}{6} \{S_2(t_f) + 4S_2(0.5t_a + t_f) + S_2(t_a + t_f)\}.$$

4 Information Time

In the monitoring process of a sequential trial, the statistically appropriate measure of how far a trial has progressed is the amount of statistical information accumulated, which is often measured by information time. Thus, it is important to have an accurate estimate of the

information time at each interim analysis. However, the information time $t^* = V(t)/V(\tau)$ derived from the log-rank test cannot be used because $V(\tau)$ is unknown until the end of the study. The traditional information time defined as the ratio of the number of events to be observed at an interim analysis to the total number in the entire trial [21] is not appropriate for HCTs as shown by our simulation study, and this deviance may be partly because of that the total number of events of the historical controls is known and remains the same at each stage of the interim analysis.

Lan and Zucker [30] proposed an information for the log-rank test at calendar time t given by

$$I(t) = \left(\frac{1}{N_1(t)} + \frac{1}{N_2(t)} \right)^{-1} \frac{d(t)}{N(t)}, \quad (10)$$

where $d(t)$ is the expected total number of events in two groups at time t under the null hypothesis, $N_i(t)$ is the sample size at time t in the i^{th} group, and $N(t) = N_1(t) + N_2(t)$. Only when a subject has had an event by time t is the information of one's survival complete. Thus, we can modify $I(t)$ in (10) as follows:

$$I(t) \simeq \left(\frac{1}{D_1(t)} + \frac{1}{D_2(t)} \right)^{-1}, \quad (11)$$

where $D_i(t)$ is the expected number of events by time t in the i^{th} group, where $i = 1, 2$. Therefore, the information time $t^* = V(t)/V(\tau) \simeq I(t)/I(\tau)$ at calendar time t can be approximated by

$$t_D^* = D(t)/D(\tau), \quad (12)$$

where $D(t) = (D_1^{-1}(t) + D_2^{-1}(t))^{-1}$ and $D(\tau) = (D_1^{-1}(\tau) + D_2^{-1}(\tau))^{-1}$

For HCTs, at calendar time t of an interim analysis, the data accumulated up to time t from the experimental group are compared to all the data from the historical controls. Thus, D_1 is not dependent on the time of an interim analysis. Therefore, we can calculate the information time at the planned calendar time t for the interim analysis by the formula

$t_D^* = D(t)/D(\tau)$, where $D(t) = (D_1^{-1}(t) + D_2^{-1}(t))^{-1}$ and $D(\tau) = (D_1^{-1}(\tau) + D_2^{-1}(\tau))^{-1}$ which can be rewritten as

$$t_D^* = \frac{(1+\xi)I}{1+\xi I}, \quad (13)$$

where $I = D_2(t)/D_2(\tau)$ is the information time for the experimental group and $\xi = D_2(\tau)/D_1$ is the ratio of the projected number of events for the experimental group to the observed

number of events for the historical controls. The information time t_D^* combining the information from the experimental group and the historical control group is called the transformed information time [19]. Because D_1 is known from the historical control data, the information time t_D^* can be estimated by letting $D_2(t) = N_2 p_2(t)$, which is the expected number of events in the experimental group up to time t , where $p_2(t)$ can be calculated by the following equation under the uniform accrual distribution:

$$p_2(t) = \frac{1}{t_a} \int_0^{t \wedge t_a} \{1 - S_2(t - u)\} du, \quad (14)$$

where $S_2(t) = [S_1(t)]^{1/\delta}$.

If the trial is designed as a maximum information trial, that is, the trial will continue until a pre-specified number of events $D_2(\tau)$ are observed for the experimental group if no early interim stopping has occurred, then the information time at the k^{th} look planned at the number of events D_{2k} for the experimental group can be calculated by

$$t_k^* = \frac{(1+\xi) I_k}{1+\xi I_k}, \quad (15)$$

where $I_k = D_{2k}/D_2(\tau)$ and $\xi = D_2(\tau)/D_1$. Given t_k^* and using equation (15) in reverse, we can obtain the calendar time t_k at the k^{th} stage corresponding to the information time t_k^* . As shown in a later example, the calendar time t_k of the k^{th} stage is indicative for calculating the sequential test statistic $L(t_k)$ by (4).

5 Group Sequential Procedure

In this section, we will apply an SCPRT procedure [22] to the log-rank test $L(t)$ to derive the sequential boundaries for HCTs. The SCPRT has two unique features: (i) the maximum sample size of the sequential test is not greater than the size of the reference fixed sample test; and (ii) the probability of discordance, or the probability that the conclusion of the sequential test would be reversed if the experiment were not stopped according to the stopping rule but continued to the planned end, can be controlled to an arbitrarily small level [19]. The SCPRT sequential design based on information time has been implemented in a user-friendly software, SCPRTinfWin, developed by Xiong [31]. Thus, study design using the SCPRT procedure is attractive and simple.

We apply the SCPRT to the test statistic $B_{t^*} = U(t)/V^{1/2}(\tau) \sim N(\theta t^*, t^*)$, which is a Brownian motion in information time $t^* = V(t)/V(\tau)$ on $[0, 1]$, and the drift parameter $\theta = \log(\delta) V^{1/2}(\tau)$. Note that, for convenience, we use the notation t^* to represent a general information time in this section. Therefore, the conditional density $f(B_{t^*} | B_1)$ is the normal density of $N(B_1 t^*, (1-t^*)t^*)$. Let $s_0 = z_{1-\alpha}$ be the critical value of B_1 to reject the null for the fixed sample test. Then the conditional maximum likelihood ratio for the stochastic process on information time t^* is as follows [32]:

$$L(t^*, B_{t^*} | z_{1-\alpha}) = \frac{\max_{\{s > s_0\}} f(B_{t^*} | B_1 = s)}{\max_{\{s \leq s_0\}} f(B_{t^*} | B_1 = s)}.$$

L can be simplified by steps in equation (1) in [32] and steps in equations (5), (6) and (7) in [22]. Then, taking the logarithm, the log-likelihood ratio can be simplified as

$$\log(L(t^*, B_{t^*} | z_{1-\alpha})) = \pm \frac{(B_{t^*} - z_{1-\alpha} t^*)^2}{2(1 - t^*) t^*} \quad (16)$$

which has a positive sign if $B_{t^*} > z_{1-\alpha} t^*$ and a negative sign if $B_{t^*} < z_{1-\alpha} t^*$. Suppose k^{th} interim looks are planned at calendar time t_k , $k = 1, \dots, K$, and the sequential test should stop when the likelihood ratio is too large or too small (i.e., the negative log likelihood ratio is too large). For given $a > 0$, with positive sign in (16), the inequality $\log(L(t_k^*, B_{t_k^*} | z_{1-\alpha})) \geq a$ is equivalent to $B_{t_k^*} - z_{1-\alpha} t_k^* \geq \{2at_k^*(1 - t_k^*)\}^{1/2}$, or to $B_{t_k^*} \geq b_k$ with negative sign in (16), the inequality $\log(L(t_k^*, B_{t_k^*} | z_{1-\alpha})) \leq -a$ is equivalent to $z_{1-\alpha} t_k^* - B_{t_k^*} \geq \{2at_k^*(1 - t_k^*)\}^{1/2}$, or to $B_{t_k^*} \leq a_k$, where

$$a_k = z_{1-\alpha} t_k^* - \{2at_k^*(1 - t_k^*)\}^{1/2}; \quad b_k = z_{1-\alpha} t_k^* + \{2at_k^*(1 - t_k^*)\}^{1/2}, \quad (17)$$

for $k = 1, \dots, K$, where $t_k^* = V(t_k)/V(\tau)$ is the information time at the k^{th} look at calendar time t_k . The a in (17) is the boundary coefficient, and it is crucial to choose an appropriate a for the design such that the probability of the conclusion by the sequential test being reversed by the test at the planned end is small but not unnecessarily so. The larger the value of a , the smaller the discordance probability and the further apart the upper and lower boundaries, making it harder for the sample path to reach the boundaries and stop early and resulting in larger expected sample sizes. Thus, an appropriate a can be determined by choosing an appropriate maximum conditional discordance probability ρ (or a maximum discordance probability ρ_{max}) from Table 1 in [32]; and usually, $\rho = 0.02$ (or $\rho_{max} = 0.0054$ for a balanced sequential design) is recommended. The nominal critical p -values for testing H_0 are

$$P_{a_k} = 1 - \Phi(a_k / \sqrt{t_k^*}); \quad P_{b_k} = 1 - \Phi(b_k / \sqrt{t_k^*}) \quad (18)$$

The observed p -value at the k^{th} look is

$$P_{B_{t_k^*}} = 1 - \Phi(B_{t_k^*} / \sqrt{t_k^*}).$$

The stopping rule for monitoring the trial can be executed by stopping the trial when, for the first time, $P_{B_{t_k^*}} \geq P_{a_k}$ (accept H_0 and stop for futility) or $P_{B_{t_k^*}} \geq P_{b_k}$ (reject H_0 and stop for efficacy). Since $L(t_k)$ by (4) has the same asymptotic distribution under the null hypothesis as $B_{t_k^*} / \sqrt{t_k^*}$, the lower and upper boundaries for $L(t_k)$ at the k^{th} look are given by

$$\bar{a}_k = a_k / \sqrt{t_k^*} \text{ and } \bar{b}_k = b_k / \sqrt{t_k^*}, \quad (19)$$

respectively, and the observed p -value at the k^{th} stage can be calculated from the test statistic $L(t_k)$ by applying all observations up to stage k . In practice, the distribution under the null is estimated from historical data by a parametric survival distribution or a spline version of the survival distribution. If an interim look is planned at a given calendar time t , then one can use the information time $t_D^* = D(t)/D(\tau)$ defined in (13) to replace the exact information time $t^* = V(t)/V(\tau)$ which is unknown because $V(\tau)$ is unknown yet at time t . If an interim look is planned at the time when the cumulative number of events reaches to a given number, then one can calculate the information time at the k^{th} interim look by

$t_k^* = (1+\xi) I_k / (1+\xi I_k)$, as defined in equation (15) for a maximum information trial where the trial is planned to be stopped after a planned number of events are observed in two groups at the final stage.

Remarks: For HCTs, the purpose to stop the trials is often for futility; thus, a one-sided sequential boundary (lower boundary) should be used in this situation. The SCPRT procedure has the feature that part or all of its upper and lower sequential boundaries can be omitted with very little effect on its significance level and power of testing; only the expected sample size of the sequential design may be substantially changed. Therefore, one can simply omit the upper boundary to obtain a one-sided lower boundary for futility monitoring. However, in trials of a new drug, it is also important to know as early as possible if the trial is highly promising; thus, the upper boundary can serve as an efficacy report instead of stopping the trial.

6 Example

In this section, we illustrate a survival HCT design. Between January 1974 and May 1984, the Mayo Clinic conducted a double-blind randomized trial on treating primary biliary cirrhosis of the liver (PBC), comparing the drug D-penicillamine (DPCA) with a placebo ([33], page 2). PBC is a rare but fatal chronic liver disease of unknown cause, with a prevalence of approximately 50 cases per million in the population. The primary pathologic event appears to be the destruction of the interlobular bile ducts, which may be mediated by immunologic mechanisms. Of 158 patients treated with DPCA, a total of 65 died. The median survival time was 9 years. Suppose a new treatment is now available and investigators want to design a trial using the Mayo Clinic patients treated with DPCA as the historical control group. The survival distribution of the DPCA data was estimated by a Kaplan-Meier curve and a logspline version of the survival distribution was fitted using the

R function **oldlogspline** (Figure 1). Assume that the survival distribution of the experimental group $S_2(t)$ satisfies the proportional hazards model

$$S_1(t) = [S_2(t)]^\delta,$$

where δ is the hazard ratio of the historical controls versus experimental group. The study aim is to test the following hypothesis:

$$H_0: \delta \leq 1 \text{ vs } H_1: \delta > 1,$$

with significance level $\alpha = 0.05$ and power of $1 - \beta = 90\%$ to detect an alternative hazard ratio $\delta = 1.75$. Thus, the required number of deaths by (7) is

$$D_2(\tau) = \left\{ \frac{[\log(1.75)]^2}{(1.645 + 1.282)^2} - 65^{-1} \right\}^{-1} = 48.$$

To calculate the sample size, we assume that subjects are recruited with a uniform distribution over the accrual period $t_a = 5$ years and followed for a period of $t_f = 4$ years, with no subjects being lost to follow-up. Then the administrative censoring distribution incurred by staggered entry is a uniform distribution over the interval $[t_f, t_a + t_f]$. The probability of death for a patient in the experimental group during the trial can be calculated by numerical integration using the spline version of the survival distribution $S_1(t)$ as follows:

$$p_2(\tau) = 1 - \frac{1}{t_a} \int_{t_f}^{t_a+t_f} [S_1(t)]^{1/1.75} dt = 0.239.$$

Thus, the trial requires enrollment of $N_2 = D_2(\tau)/p_2(\tau) = 201$ patients. The empirical power and type I error are 89.6% and 0.049, respectively, based on 100,000 simulated trials from the fitted spline distribution. Suppose three interim looks are planned at 20, 30, and 48 events. Then we have $\xi = D_2(\tau)/D_1 = 48/65 = 0.378$, $I_1 = D_{21}/D_2(\tau) = 20/48 = 0.417$, $I_2 = D_{22}/D_2(\tau) = 30/48 = 0.625$, and $I_3 = D_{23}/D_2(\tau) = 1$. Thus, the transformed information times calculated by the formula

$$t_k^* = \frac{(1+\xi) I_k}{1+\xi I_k}$$

are given as $t^* = (t_1^*, t_2^*, t_3^*) = (0.496, 0.697, 1)$. If we set the type I error to $\alpha = 0.05$, power of $1 - \beta = 90\%$, number of interim looks $K = 3$ (including the final look), and information time $t^* = (0.496, 0.697, 1)$ at each stage in the software SCPRTinfWin, and given a maximum conditional discordance probability $\rho = 0.02$, it calculates the boundary coefficient $a = 2.596$, the lower boundaries $(a_1, a_2, a_3) = (-0.323, 0.099, 1.645)$ and the upper boundaries $(b_1, b_2, b_3) = (1.955, 2.194, 1.645)$ by (17), the nominal critical p -values for the lower boundaries $(P_{a_1}, P_{a_2}, P_{a_3}) = (0.6769, 0.4526, 0.05)$, and those for the upper boundaries

$(P_{b_1}, P_{b_2}, P_{b_3}) = (0.0028, 0.0043, 0.05)$. To monitor the trial at the k^{th} interim look, the survival data collected up to the k^{th} interim look from the experimental group is combined with all the data from the historical controls to calculate the sequential test statistic $L_k = L(t_k)$ by (4), where t_k is the calendar time at the k^{th} stage that is either being observed when the cumulative number of events has just reached the predetermined number, or is being calculated as described at the end of Section 4. The observed p -values are

$$P_{L_k} = 1 - \Phi(L_k), \quad k=1, 2, 3.$$

At the k^{th} stage, we stop the trial for futility if $P_{L_k} \geq P_{a_k}$, and we stop the trial for efficacy (or report efficacy) if $P_{L_k} \leq P_{b_k}$. The operating characteristics for the proposed three-stage group sequential design are shown in Table 1.

7 Conclusion

In this paper, the sample size calculation for survival HCTs is derived. In addition, a transformed information time that is not only based on the information time of the experimental group but also uses the information from the historical controls is derived for trial monitoring. It is simple and convenient to use the transformed information time to derive the sequential boundaries for the HCTs based on the SCPRT procedure. With this monitoring procedure, data from the experimental group are sequentially collected and compared with all data from the historical controls. This allows investigators to monitor the trial at any calendar time during enrollment and follow-up or at the time a pre-specified number of events is reached, which has been predetermined as an interim look. A user-friendly software, SCPRTinfWin [31], is available for HCT design. Therefore, the proposed group sequential procedure is simple to use. For the purpose of study design, we need only the number of events from the historical control data to calculate the sample size. However, for trial monitoring and final data analysis, we need the full data from the historical control group in order to calculate the sequential test statistic. In practice, the historical control data are often available from previous trials conducted by the same institution or sponsor. If no such historical control data is available, then we need to extract the relevant data from the published literature. Recently, Guyot et al. [34] have proposed a method for reconstructing the survival data from published Kaplan-Meier survival curves. Thus, HCTs are also feasible by using historical control data from the published literature.

In a HCT trial, the distributions of covariates of patients may differ between the historical control group and the experimental group, and this difference may lead to potential bias in the result of testing. One possible practical solution to prevent this bias is that we can first identify influential covariates of treatment effect through analysis of historical data and then enroll the experimental group according to the stratifications of the covariates identified from the historical control group. For example in the Example in Section 6, assuming the gender had been identified as the influential factor of survival in the historical control group and the ratio of male to female is 1:3, then we can enroll patients for the experimental group sequentially with a gender allocation in this proportion. In this way, the influential variates would be similarly distributed in both historical and experimental groups.

If this practical way of matched enrollment mentioned above is difficult or impossible to be carried out in the trial, then one may use the Cox regression model to adjust for confounding factors and the score test based on the partial likelihood can be used for testing. This score test has asymptotically the same sequential distribution as the log-rank test with no covariates [35], however, the power of two procedures could be different. The use of the log-rank test for the testing could result in a loss of power when a covariate effect is present. It will be a topic of future research to develop a group sequential procedure based on the score test derived from the partial likelihood for the historical control trial design when covariates need to be taken account.

There are several ways to adjust the treatment comparison after the HCT trial. Propensity scores have been used to reduce bias of such comparison in observational studies. The propensity score is the probability receiving treatment given covariates, which can be estimated by a logistic regression model. One can then include the propensity score as a covariate in the Cox regression analysis. Recently, Xie and Liu [37] proposed a weighted log-rank test which adjusts for confounding by using propensity scores to construct weights for individual observations. Applying this method to the HCT trial, one can pool the patients of experimental group and historical control group together to establish a logistic model predicting status of groups by covariates, then using this model to calculate the propensity score for each patient in the two groups where the propensity score is the probability of being in the experimental group for given covariates. The test procedure can be obtained by adding a weight to the item of $U(t)$ in equation (3) where the weight is the inverse of propensity score. The treatment effect can then be tested by using this inverse propensity score weighted log-rank test, which has been implemented in SAS version 12.1.

Acknowledgments

The author acknowledges anonymous reviewers and associate editor for their valuable comments that improved an earlier version of the paper. This work was supported in part by National Cancer Institute (NCI) support grant P30CA021765 and ALSAC.

References

1. Pocock S. Randomized versus historical controls: a compromise solution. *Proceedings of the International Biometric Conference*. 1976; 9:245-260.
2. Farewell VT, D'Angio GJ. A simulated study of historical controls using real data. *Biometrics*. 1981; 37:169–176. [PubMed: 6264985]
3. Gehan EA. Design of controlled clinical trials: use of historical controls. *Cancer Treatment Reports*. 1982; 66:1089–1093. [PubMed: 6211230]
4. Fleming TR. Historical controls, data banks, and randomized trials in clinical research: a review. *Cancer Treatment Reports*. 1982; 66:1101–1105. [PubMed: 7083214]
5. Green SG, Byar DP. Using observational data from registries to compare treatments: the fallacy of omnimetrics. *Statistics in Medicine*. 1984; 3:361–370. [PubMed: 6396792]
6. Gehan EA, Freireich EJ. A rational basis for use of historical controls. *Cancer Clinical Trials*. 1981; 8:430–436.
7. 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; 44:80–86. [PubMed: 19159722]

8. Abe T, Kakemura T, Fujinuma S, Maetani I. Successful outcomes of EMRL with 3D-EUS for rectal carcinoids compared with historical controls. *World Journal of Gastroenterology*. 2008; 14:4054–4058. [PubMed: 18609690]
9. 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*. 2008; 12:1–8.
10. 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; 29:991–996. [PubMed: 18310232]
11. 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; 21:53–60.
12. 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; 46:431–436.
13. 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]
14. Makuch RW, Simon RM. Sample size considerations for non-randomized comparative studies. *Journal of Clinical Epidemiology*. 1980; 33:175–181.
15. Dixon DO, Simon R. Sample size considerations for studies comparing survival curves using historical controls. *Journal of Clinical Epidemiology*. 1988; 41:1209–1213. [PubMed: 3062140]
16. Emrich LJ. Required duration and power duration determinations for historically controlled studies of survival times. *Statistics in Medicine*. 1989; 8:153–160. [PubMed: 2704897]
17. Korn EL, Freidlin B. Conditional power calculations for clinical trials with historical controls. *Statistics in Medicine*. 2006; 25:2922–2931. [PubMed: 16479548]
18. Chang MN, Shuster JJ, Kepner JL. Group sequential designs for phase II trials with historical controls. *Controlled Clinical Trials*. 1999; 20:353–364. [PubMed: 10440562]
19. Xiong X, Tan M, Boyett J. A sequential procedure for monitoring clinical trials against historical controls. *Statistics in Medicine*. 2007; 26:1497–1511. [PubMed: 16900551]
20. Zhang S, Gao J, Ahn C. Calculating sample size in trials using historical controls. *Clinical Trials*. 2010; 7:343–353. [PubMed: 20573638]
21. Lan KKG, Lachin JM. Implementation of group sequential logrank tests in a maximum duration trial. *Biometrics*. 1990; 46:759–770. [PubMed: 2242413]
22. Xiong X. A class of sequential conditional probability ratio tests. *Journal of the American Statistical Association*. 1995; 15:1463–1473.
23. Tsiatis AA. Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association*. 1982; 77:855–861.
24. Sellke T, Siegmund D. Sequential analysis of the proportional hazards model. *Biometrika*. 1983; 79:315–326.
25. Slud EV. Sequential linear rank tests for two-sample censored survival data. *Annals of Statistics*. 1984; 12:551–571.
26. Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*. 1981; 68:316–319.
27. Hsieh FY. Comparing sample size formulae for trials with unbalanced allocation using the logrank test. *Statistics in Medicine*. 1992; 11:1091–1098. [PubMed: 1496196]
28. Rubinstein LV, Gail MH, Santner TJ. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of chronic Diseases*. 1981; 34:469–479. [PubMed: 7276137]
29. Kooperberg C, Stone CJ. Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*. 1992; 1:301–328.

30. Lan KKG, Zucker DM. Sequential monitoring of clinical trials: the role of information and Brownian motion. *Statistics in Medicine*. 1993; 12:753–765. [PubMed: 8516592]
31. Xiong, X. SCPRTinfWin: sequential conditional probability ratio tests on information time. 2007. <http://www.stjude-research.org/depts/biostats/scprt>
32. Xiong X, Tan M, Boyett J. Sequential conditional probability ratio tests for normalized test statistic on information time. *Biometrics*. 2003; 59:624–631. [PubMed: 14601763]
33. Fleming, TR., Harrington, DP. Counting processes and survival analysis. John Wiley and Sons; New York: 1991.
34. Guyot P, Ades AE, Ouwens M, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology*. 2012; 12:1–13. [PubMed: 22214542]
35. Tsiatis AA, Rosner GL, Titchler DL. Group sequential tests with censored survival data adjusting for covariates. *Biometrika*. 1985; 72:365–373.
36. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55.
37. Xie J, Liu C. Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine*. 2005; 24:3089–3110. [PubMed: 16189810]

Figure 1.

footnote: The step function is the Kaplan-Meier survival curve and the dotted lines are the 95% confidence boundaries. The solid curve is the fitted logspline survival distribution.

Table 1

Operating characteristics of the three-stage SCPRT design for the sequential test statistic $L(t)$ based on 100,000 simulation runs under the logspline distribution with uniform censoring distribution over $[t_F, t_a + t_F]$, nominal type I error of 0.05, and power of 90% for the example in Section 6, where $t_a = 5$ and $t_F = 4$.

At k^{th} interim look	$k = 1$	$k = 2$	$k = 3$	Total
Type I error				
Empirical of $L(t_k)$	0.0026	0.0029	0.0444	0.0499
Nominal	0.0028	0.0030	0.0448	0.0506
Power				
Empirical of $L(t_k)$	0.2304	0.2185	0.4474	0.8963
Nominal	0.2373	0.2085	0.4536	0.8994
Probability of stopping under null				
Empirical of $L(t_k)$	0.3261	0.2390	0.4349	1.0000
Nominal	0.3258	0.2450	0.4293	1.0000
Probability of stopping under alternative				
Empirical of $L(t_k)$	0.2352	0.2244	0.5404	1.0000
Nominal	0.2432	0.2154	0.5416	1.0000