Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Wu and Xiong, J Biomet Biostat 2014, 5:4
DOI: 10.472/2155-6180.1000209

Research Article Open Access

Group Sequential Survival Trial Designs Against Historical Controls under the Weibull Model

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, two parametric sequential tests are proposed for historical control trial designs under the Weibull model. The proposed tests are asymptotically normal with properties of Brownian motion. The sample size formulas and information times are derived for both tests. A multi-stage sequential procedure based on sequential conditional probability ratio test methodology is proposed for monitoring clinical trials against historical controls.

Keywords: Brownian motion; Group sequential trial; Historical control; Information time; Sample size; Time-to-event; Weibull distribution

Introduction

Randomized clinical trials are the gold standard for comparing a new therapy to a standard treatment. However, when randomization is not feasible because of ethical concerns, patient preference, or regulatory acceptability, comparing data from patients receiving a new therapy to those from patients previously treated by standard treatment (historical control) is an alternative. If patients enrolled in the current trial are similar to those in the historical study, clinical trials with a historical control improve the reliability of testing results of single-arm phase II trials by including the variation of the null parameter, which is usually estimated from historical data. Compared with randomized phase III trials, clinical trials with a historical control require a much smaller sample size, and are therefore easier to conduct and save time and patient resources [1].

Despite the practical and statistical issues associated with historical control trials [2-6], they have been appropriately used in many clinical practices. Sample size calculations to design such trials have been discussed by Makuch and Simon [7] for binary endpoints and by Dixon and Simon [8] and Emrich [9] for exponential survival endpoints. These methods have been widely used in oncology trial designs. However, Korn and Freidlin [10] reported that these popular methods do not preserve the power and type I error when considering the uncertainty in the historical control outcome data. Recently, several studies have discussed sample size calculations for historical control trials by taking into account the uncertainty in historical control outcome data [11-13].

Clinical trials with historical controls are often monitored by pre-planned interim analyses to stop accrual if patients in the current trial have poorer outcomes than those in the historical control. The monitoring of clinical trials with historical controls poses a statistical problem of comparing two outcomes in a situation wherein data from the current study are sequentially collected and compared with all data from historical controls at each interim analysis. Few studies have discussed the monitoring of clinical trials against historical controls. For example, Chang et al. [11] proposed a two-stage design for binary outcome and Xiong et al. [12] developed a multistage group sequential procedure for monitoring historical control trials with binary, continuous, and survival endpoints.

In this study, we propose a multistage group sequential procedure to de-sign survival trials against historical controls under the Weibull model. In Section 2, two sequential parametric tests are proposed for the trial design under the Weibull model. In Section 3, formulas for the number of events required for the current study are derived. In Section 4, a multistage group sequential procedure based on the sequential conditional probability ratio test (SCPRT) by Xiong [1] is proposed. In Section 5, simulation studies to calculate the empirical power and type I error of the proposed tests are described. In Section 6, an example is given to illustrate the proposed methods. The discussion and concluding remarks are given in Section 7.

Sequential Test Statistics

Two parametric sequential test statistics are discussed in this section to provide group sequential design of survival trials against historical controls under the Weibull model. Assume that the failure time variable T_j of a subject from the j^{th} group follows the Weibull distribution with a common shape parameter κ and a scale parameter ρ_j, where j=1 for the historical control group and j=2 for the current study group. That is, T_j has survival distribution function

$$S_j(t) = e^{-(\rho_j t)^\kappa}$$

and hazard function

$$h_j(t) = \kappa \rho_j^\kappa t^{\kappa-1}.$$

The shape parameter κ indicates the degree of acceleration (κ>1), constant (κ=1), or deceleration (κ<1) of the hazard over time. In a cancer trial, the median survival time is an intuitive endpoint for clinicians. The median survival time of the j^{th} group for the Weibull distribution can be calculated as $m_j = \rho_j^{-1}\{log(2)\}^{1/\kappa}$. Therefore, the Weibull survival distribution can be expressed as

$$S_j(t) = e^{-log(2)\left(\frac{t}{m_j}\right)^\kappa}, j = 1,2.$$

The one-sided hypotheses of a historical control trial defined by

***Corresponding author:** Jianrong Wu, Department of Biostatistics, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, USA, Tel: (901) 495-2850; Fax: (901) 495-4585; E-mail: jianrong.wu@stjude.org

Received July 04, 2014; **Accepted** August 04, 2014; **Published** August 11, 2014

Citation: Wu J, Xiong X (2014) Group Sequential Survival Trial Designs Against Historical Controls under the Weibull Model. J Biomet Biostat 5: 209. doi:10.4172/2155-6180.1000209

median survival times can be expressed as a two-sample test of the following:

$$H_0 : m_2 \le m_1 \; vs. H_1 : m_2 \le m_1.$$

For notational convenience, we convert the scale parameter ρ_j to a hazard parameter $\lambda_j = \rho_j^{\kappa} = log(2)/m_j^{\kappa}$. Then the survival distribution is $S_j(t) = e^{-\lambda_j t^{\kappa}}$ with hazard function $h_j(t) = \kappa\lambda_j t^{\kappa-1}$, in which κ is taken as a known constant. In this case, the above hypotheses on median survival times are equivalent to

$$H_0^* : \delta \le 1 \; vs. \; H_1^* : \delta > 1,$$

where the hazards ratio $\delta = \lambda_1/\lambda_2 = (m_2/m_1)^{\kappa}$.

Now, suppose there are n_1 subjects on the historical control group and let T_{i1} and C_{i1} denote, respectively, the failure time and censoring time of the i^{th} subject of the historical control group. Further assume that during the accrual phase of the current trial, n_2 subjects are enrolled in the study, and let T_{i2} and C_{i2} denote, respectively, the failure time and censoring time of the i^{th} subject of the current study group, with both being measured from the time of study entry Y_{i2}. We assume that the failure time T_{ij} is independent of the censoring time C_{ij} and entry time Y_{i2}, and $\{(T_{ij}, C_{ij}, Y_{i2}); i=1,\cdots, n_j\}$ are independent and identically distributed. When the current study data are examined at calendar time $t \le \tau$, where τ is the study duration, we observe the time to failure $X_{i1} = T_{i1} \wedge C_{i1}$ and the failure indicator $\Delta_{i1} = I(T_{i1} \le C_{i1}), i = 1,\cdots, n_1$ for the historical control, and the time to failure $X_{i2}(t) = (T_{i2} \wedge C_{i2}) \wedge (t - Y_{i2})^+$ and the failure indicator $\Delta_{i2}(t) = I(T_{i2} \le C_{i2} \wedge (t - Y_{i2})^+), i = 1,\cdots, n_2$ for current study group up to time t. We assume that survival data remain the same (no further follow-up) for the historical control during the process of the current trial, whereas survival data are updated for the current study from one look to the next in the trial. On the basis of the observed data $\{X_{i1}, \Delta_{i1}, X_{i2}(t), \Delta_{i2}(t)\}$ at interim look time t, the observed likelihood function is proportional to ([14], Chapter 3)

$$L(\lambda_1, \lambda_2; t) = \lambda_1^{d_1} \lambda_2^{d_2(t)} e^{-\lambda_1 U_1 - \lambda_2 U_2(t)},$$

where $d_1 = \sum_{i=1}^{n_1} \Delta_{i1}^{\kappa}$ is the total number of events of the historical control; $U_1 = \sum_{i=1}^{n_1} X_{i1}^{\kappa}$ is the cumulative follow-up time of historical controls penalized by the Weibull shape parameter κ; $d_2(t) = \sum_{i=1}^{n_2} \Delta_{i2}(t)$ is the total number of events observed in the current group by time t; and $U_2(t) = \sum_{i=1}^{n_2} X_{i2}^{\kappa}(t)$. The maximum likelihood estimates of λ_1 and λ_2 can be derived as

$$\hat{\lambda}_1 = d_1/U_1 \text{ and } \hat{\lambda}_2 = d_2(t)/U_2(t)$$

with variances $\hat{\lambda}_1^2/d_1$ and $\hat{\lambda}_2^2(t)/d_2(t)$, respectively. Therefore, under the null hypothesis, the Wald statistic of the log-hazard ratio $log(\delta)$ at calendar time t is given by

$$Z(t) = log\{d_1 U_2(t)/d_2(t) U_1\}(d_1^{-1} + d_2^{-1}(t))^{-1/2}, \quad (1)$$

which has approximately a standard normal distribution. To derive the group sequential design, let

$$U(t) = log\{d_1 U_2(t)/d_2(t) U_1\}(d_1^{-1} + d_2^{-1}(t))^{-1}; \quad (2)$$

then under the alternative of $\delta = \lambda_1/\lambda_2 > 1$, the statistic $U(t) = Z(t)/(d_1^{-1} + d_2^{-1}(t))^{1/2}$ is approximately normal with mean $log(\delta)V(t)$ and variance $V(t)$ and has an independent increment structure, where $V(t) = (d_1^{-1} + d_2^{-1}(t))^{-1}$. The above results can be derived from Tsiatis et al. [15], who reported similar results for general parametric survival models. Because

$$V(t) \sim D(t) = (D_1^{-1} + D_2^{-1}(t))^{-1}, \quad (3)$$

where D_1 is the total number of events in the historical control and $D_2(t) = n_2 p_2(t) = n_2 P(\Delta_{12}(t) = 1)$ is the number of events in the current study up to time t. Thus, $B_{t^*} = U(t)/D^{1/2}(\tau) \sim N(\theta t^*, t^*)$ is approximately a Brownian motion with drift parameter $\theta = log(\delta) D^{1/2}(\tau)$ and information time $t^* = D(t)/D(\tau)$, where $D(\tau)$ is the value of $D(t)$ at $t=\tau$.

Sprott [16] showed that the distribution of $\hat{\phi}_1 = \hat{\lambda}_1^{1/3}$ and $\hat{\phi}_2(t) = \hat{\lambda}_2^{1/3}(t)$ in small samples is much more closely approximated by a normal distribution. Then $\hat{\phi}_1 = \hat{\lambda}_1^{1/3}$ and $\hat{\phi}_2(t) = \hat{\lambda}_2^{1/3}(t)$ are approximately normal with mean $\phi_j = \lambda_j^{1/3}, j = 1, 2$ and variance estimate $\hat{\phi}_1^2/(9d_1)$ and $\hat{\phi}_2(t)^2/(9d_2(t))$ [17]. Therefore, the test statistic

$$S(t) = \frac{\hat{\phi}_1 - \hat{\phi}_2(t)}{\{\hat{\phi}_1^2/(9d_1) + \hat{\phi}_2^2/(9d_2(t))\}^{1/2}},$$

is an approximately standard normal distribution under the null hypothesis.

Let

$$U(t) = \frac{3(\hat{\phi}_1 - \hat{\phi}_2(t))}{\hat{\phi}_2(t)((\hat{\phi}_1^2/\hat{\phi}_2^2(t))/d_1 + 1/d_2(t))},$$

then under the alternative, the statistic $U(t) = S(t)/((\hat{\phi}_1^2/\hat{\phi}_2^2(t))d_1^{-1} + d_2^{-1}(t))^{1/2}$ is approximately normal with mean $3(\delta^{1/3} - 1)V(t)$ and variance $V(t)$, where $V(t) = ((\hat{\phi}_1^2/\hat{\phi}_2^2(t))d_1^{-1} + d_2^{-1}(t))^{-1}$, and $U(t)$ has an independent increment structure. Because

$$V(t) \simeq D(t) = (\delta^{2/3} D_1^{-1} + D_2^{-1}(t))^{-1}, \quad (4)$$

$B_{t^*} = U(t)/D^{1/2}(\tau) \sim N(\theta t^*, t^*)$ is approximately a Brownian motion with drift parameter $\theta = 3(\delta^{1/3} - 1)D^{1/2}(\tau)$ and information time $t^* = D(t)/D(\tau)$.

Sample Size for Fixed Sample Test

Because historical control data are obtained from previous trials, sample size n1 and total number of events D1 for the historical control group are known. Therefore, we only need to calculate the sample size for the current study for a fixed sample test at the end of the study. On the basis of the test statistic $Z(t)$ at $t=\tau$, under the null hypothesis,

$$Z(\tau) = log\{d_1 U_2(\tau)/d_2(\tau) U_1\}(d_1^{-1} + d_2^{-1}(\tau))^{-1/2}$$

has an approximately standard normal distribution. To calculate the power under the alternative $\delta = \lambda_1/\lambda_2 (> 1)$, $Z(\tau)$ is an approximately normal distribution with mean $log(\delta)D^{1/2}(\tau)$ and unit variance. Therefore, given a significance level α, the power $(1-\beta)$ of the $Z(\tau)$ test under the alternative is given by

$$1 - \beta \simeq \Phi\{log(\delta)(D_1^{-1} + D_2^{-1}(\tau))^{-1/2} - z_{1-\alpha}\},$$

where $\Phi(\cdot)$ is the standard normal distribution function and $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$. Thus, the number of events required for the current study based on the $Z(\tau)$ test can be calculated by

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

where $\delta = (m_2/m_1)^{\kappa}$ and D_1 is the total number of events observed in

historical control data. Therefore, the sample size for the current group is given by

$$n_2 = D_2(\tau) / p_2(\tau),$$

where $p_2(\tau)$ is the probability of a subject from the current group having an event during the study. Similarly, the number of events required for the current study based on the $S(\tau)$ test can be calculated by

$$D_2(\tau) = \left\{ \frac{[9(\delta^{1/3}-1)]^2}{(z_{1-\alpha}+z_{1-\beta})^2} - \delta^{2/3} D_1^{-1} \right\}^{-1}, \quad (6)$$

and the sample size is given by $n_2 = D_2(\tau) / p_2(\tau)$.

To calculate the number of subjects required for the study, we need to calculate $p_2(\tau)$, the probability of a subject in the current group having an event during the study. Typically, we assume that subjects are accrued over an accrual period of length t_a with an additional follow-up period of length t_f. A subject enters the study at time u, the entry time is uniformly distributed on $[0,t_a]$, and no subject is lost to follow-up during the study. Then the probability of a subject having an event during the study under the Weibull model can be calculated by [18]

$$p_2(\tau) = 1 - \frac{1}{t_a} \int_{t_f}^{t_a+t_f} e^{-\log(2)\left(\frac{u}{m_2}\right)^{\kappa}} du. \quad (7)$$

Therefore, given the design parameters δ (or κ), m_1,m_2, α, β, t_f and t_a, the number of subjects n_2 required for the current study can be calculated by $n_2 = D_2(\tau) / p_2(\tau)$ and using the formula in equation (5) or (6).

In designing an actual trial, given the accrual time t_a, calculating the sample size is often impractical because it may be not possible to enroll the total number of subjects as planned in the given accrual duration. It is more practical to design the study starting with the accrual rate r and then calculate the required accrual time t_a. This can be accomplished under the Weibull model assumption. First, the integration in the probability formula (7) can be simplified by approximation, using the Simpson rule

$$p_2(\tau) = 1 - \frac{1}{6}\left\{ S_2(t_f) + 4S_2(0.5t_a + t_f) + S_2(t_a + t_f) \right\}. \quad (8)$$

Then, combining the sample size formula based on equations (5) or (6) with equation (8), we can define a root function of the accrual time t_a

$$\text{root}(t_a) = rt_a - \left\{ \frac{[log(\delta)]^2}{(z_{1-\alpha}+z_{1-\beta})^2} - D_1^{-1} \right\}^{-1} / p_2(t_a + t_f). \quad (9)$$

Now the accrual time ta can be obtained by solving the root equation root(t_a)=0 numerically in Splus using the uniroot function. The total sample size required for the current study is approximately $n_2=[rt_a]^+$, where $[x]^+$ denotes the smallest integer greater than x.

Once the number of events or sample size is calculated for the fixed sample test, we can calculate the information time at the planned calendar time t for the interim analysis by $t^* = D(t) / D(\tau)$. For example, if we plan K interim analyses at calendar time t_k, k=1,...,K, then the information time at calendar time t_k can be calculated by $t_k^* = D(t_k) / D(\tau)$, where $D(t)$ is given by equations (3) and (4) for $Z(t)$ and $S(t)$, respectively. After some simplifications, the information time $t^* = D(t)/ D(\tau)$ can be rewritten as

$$t^* = \frac{(1+R)I}{1+RI}, \quad (10)$$

where $I = D_2(t)/ D_2(\tau)$ is the information time for the current study and $R=D_2(\tau)/D_1$ is the ratio of the number of events of the current study to the historical control for the $Z(t)$ test, and $R = \delta^{2/3} D_2(\tau) / D_1$ for $S(t)$ test. This is called the transformed information time [12]. Because D_1 is known from historical control data, thus, under the Weibull model, the information time t^* can be obtained by calculating $D_2(t)=n_2p_2(t)$, which is the expected number of events in the current study up to time t, where $p_2(t)=P(\Delta_{12}(t)=1)$ can be calculated as

$$p_2(t) = \frac{1}{t_a} \int_0^{t \wedge t_a} \{1 - S_2(t-u)\} du, \quad (11)$$

where $S_2(t) = e^{-\log(2)\left(\frac{u}{m_2}\right)^{\kappa}}$. When $t=\tau$, equation (11) is identical to equation (7).

For a maximum information trial where the trial continues until a pre- specified number of events $D_2(\tau)$ observed for the current study, the information time at the k^{th} look planned at number of events D_{2k} for the current study can be calculated by $t_k^* = (1+R)I_k / (1+DI_k)$, where $I_k=D_{2k}/D_2(\tau)$, $R=D_2(\tau)/D_1$ and $R = \delta^{2/3} D_2(\tau) / D_1$ for $Z(t)$ and $S(t)$, respectively.

Group Sequential Procedure

In this section, we will apply an SCPRT procedure [1] to the test statistics $Z(t)$ and $S(t)$. The SCPRT has two unique features: (1) the maximum sample size of the sequential test is not greater than the size of the reference fixed sample test; and (2) 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 [12]. Furthermore, the power function of the SCPRT is virtually the same as that of the fixed sample test [1]. The SCPRT boundaries derived in our study have analytical solutions. All these features make the SCPRT attractive and simple to use.

Now we apply the SCPRT to the test statistic $B_{t^*} = U(t) / D^{1/2}(\tau) \sim N(\theta t^*, t^*)$, which is a Brownian motion in information time $t^* = D(t) / D(\tau)$ on [0, 1], and drift parameter $\theta = log(\delta) D^{1/2}(\tau)$ for $Z(t)$ and $\theta = 3(\delta^{1/3}-1) D^{1/2}(\tau)$. Therefore, the conditional density $f(B_{t^*} \mid B_1)$ 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^* (see, [1,19]) is

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

Taking the logarithm, the log-likelihood ratio can be simplified as

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

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,...,K. Then on the basis of the SCPRT procedure presented above, the lower and upper boundaries for $B_{t_k^*}$ at the k^{th} look are given by

$$a_k = z_{1-\alpha} t_k^* - \left\{ 2at_k^* \left(1 - t_k^*\right) \right\}^{1/2}; b_k = z_{1-\alpha} t_k^* + \left\{ 2at_k^* \left(1 - t_k^*\right) \right\}^{1/2}, \quad (12)$$

for k=1,...,K, where $t_k^* - D(t_k) / D(\tau)$ is the information time at the k^{th} look at calendar time t_k. The a in (12) is the boundary coefficient, and it is crucial to choose an appropriate a for the design such that the

probability of conclusion by the sequential test being reversed by the test at the planned end is small but not unnecessarily too small. The larger the a, the smaller is the discordance probability and the wider apart are the upper and lower boundaries, making it harder for the sample path to reach boundaries and stop early and resulting in larger expected sample sizes. Thus, an appropriate a can be determined by choosing an appropriate discordance probability [1,19]. The nominal critical p-values for testing H_0 are

$$P_{a_k} = 1 - \Phi(a_k / \sqrt{t_k^*}); P_{b_k} = 1 - \Phi(b_k / \sqrt{t_k^*}). \quad (13)$$

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, $P_{B_{t_k^*}} \geq P_{b_k}$ (accept H_0 and stop for futility) or $P_{B_{t_k^*}} \leq P_{b_k}$ (reject H_0 and stop for efficacy). Because $Z(t_k)$ or $S(t_k)$ has the same asymptotic distribution as the $B_{t_k^*} / \sqrt{t_k^*}$ under the null hypothesis, the observed p-value at the k^{th} stage can be calculated from the test statistic $Z(t_k)$ or $S(t_k)$ by applying all observations up to stage k.

Simulation Studies

In this section, we conducted simulation studies to compare the power and type I error of the proposed parametric test statistics $Z(t)$ and $S(t)$ under various scenarios. In the simulations, the survival distribution of the j^{th} group was taken as $S_j(t) = e^{-\log(2)(t/m_j)^\kappa}$, which is the Weibull distribution with shape parameter κ and median survival time m_j, j=1,2, where j=1 and j=2 represent the historical control and current study, respectively. The shape parameter κ was taken as 0.5, 1, and 2.0 to reflect cases of decreasing, constant, and increasing hazard functions, respectively. We assume a median time-to-event m_1=3.4657 and a sample size n_1=140 for the historical control. The null hypothesis was set to $H_0 : m_1 = m_2$, and the hazard ratio $\delta = (m_2/m_1)^\kappa$ under the alternative was taken as 1.5-2.0. Furthermore, we assumed that subjects of the current study were recruited with a uniform distribution over the accrual period t_a=4 (years) and followed for t_f=1 (years), and no subject was lost to follow-up during the study period $\tau = t_a + t_f$=5. Therefore, a subject was censored at calendar time t if his/her event time was longer than $t-u$, where u is the time when the subject entered the current study.

In Table 1, the sample sizes required for the current study were calculated by equations (5) and (6) for test $Z(t)$ and $S(t)$, respectively. Furthermore, in each design parameter configuration, 100,000 observed samples of censored event times were generated from the Weibull distribution to calculate the test statistics under the null or alternative hypothesis. The nominal significance level and power were set to 0.05 and 80%, respectively. Two simulation studies were done. The first simulation was done to study the empirical type I error and power for the fixed sample tests. The second simulation was done to study the empirical type I error and power for a two-stage SCPRT design at calendar times t_1=3 and t_2=5. The simulated empirical type I errors and powers in various scenarios for the fixed sample tests and two-stage SCPRT tests are summarized in Tables 1 and 2, respectively. The results of the fixed sample tests showed that the $S(\tau)$ test needs a slightly larger sample size for a small δ and smaller sample size for a large δ compared with the $Z(\tau)$ test. The simulated empirical type I errors and powers were close to the nominal levels for the $S(\tau)$ test, and the $Z(\tau)$ test was somewhat overpowered for a large δ. For the two-stage design $S(t)$ had adequate empirical power and type I error

Design		δ=1.5			δ=1.6			δ=1.7		
κ	Test	$\hat{n}_2$	α	1 - β	n_2	α	1 - β	n_2	α	1 - β
0.5	$Z(\tau)$	262	0.052	0.795	152	0.051	0.804	108	0.049	0.817
	$S(\tau)$	285	0.051	0.799	149	0.05	0.802	100	0.051	0.804
	$L(\tau)$	262	0.053	0.795	152	0.051	0.804	108	0.049	0.818
1	$Z(\tau)$	305	0.053	0.793	170	0.05	0.807	118	0.049	0.813
	$S(\tau)$	344	0.052	0.799	168	0.05	0.802	111	0.052	0.806
	$L(\tau)$	305	0.053	0.793	170	0.05	0.807	118	0.049	0.813
2	$Z(\tau)$	367	0.054	0.795	191	0.053	0.803	130	0.05	0.811
	$S(\tau)$	445	0.049	0.801	195	0.051	0.8	124	0.051	0.802
	$L(\tau)$	367	0.054	0.794	191	0.053	0.802	130	0.05	0.81
Design		**δ=1.8**			**δ=1.9**			**δ=2.0**		
κ	Test	n_2	α	1 - β	n_2	α	1 - β	n_2	α	1 - β
0.5	$Z(\tau)$	84	0.049	0.823	70	0.047	0.833	60	0.046	0.84
	$S(\tau)$	75	0.05	0.806	61	0.051	0.813	51	0.051	0.815
	$L(\tau)$	84	0.049	0.823	70	0.048	0.834	60	0.047	0.842
1	$Z(\tau)$	92	0.048	0.823	75	0.046	0.828	65	0.046	0.838
	$S(\tau)$	82	0.051	0.807	66	0.05	0.812	55	0.051	0.814
	$L(\tau)$	92	0.049	0.823	75	0.047	0.828	65	0.047	0.839
2	$Z(\tau)$	99	0.048	0.817	81	0.047	0.825	69	0.047	0.833
	$S(\tau)$	90	0.051	0.805	71	0.051	0.806	59	0.051	0.812
	$L(\tau)$	99	0.048	0.815	81	0.048	0.824	69	0.048	0.832

Table 1: Sample size and simulated empirical type I error (α) and power (1-β) based on 100,000 simulation runs for the Weibull distribution for fixed sample $Z(\tau)$ test, log-rank test $L(\tau)$ and $S(\tau)$ test with a nominal type I error of 0.05 and power 80% (one-sided test).

whereas the $Z(\tau)$ test was conservative and under-powered for a large δ in the first stage. Overall, the test statistic $S(t)$ performed better than $Z(t)$ and is recommended for use in the trial design. By the way, to show if the sample size formula (5) and information time (10) developed for the $Z(t)$ test also work for the non-parametric log-rank test $L(t)$, the empirical type I errors and powers were simulated for the log-rank test too (Tables 1 and 2). The results showed that both sample size formula (5) and transformed information time (10) worked well for the log-rank test. A rigorous derivation of these results for the log-rank test will be the future research.

An Example

Between January, 1974 and May, 1984, the Mayo Clinic conduct a double- blind randomized trial in primary biliary cirrhosis (PBC), comparing the drug D-penicillamine (DPCA) with a placebo (Fleming and Harrington, 1991). PBC is a rare but fatal chromic liver disease of unknown cause, with a prevalence of about 50-cases-per-millian population. The primary pathologic event appears to be the destruction of interlobular bile ducts, which may be mediated by immunologic mechanisms. A total of 65 had died among 158 patients treated with DPCA. The median survival time was 9 years. Suppose a new treatment is now available and investigators want to design a new trial using Mayo Clinic patients treated with DPCA as the historical control group. The survival distribution of DPCA data were estimated by Kaplan-Meier method and the Weibull model. The Weibull distribution fitted the survival distribution well with shape parameter κ=1.22 and scale parameter $\rho = 11.8^{-1}$. Thus to design the study, we can assume that the failure time of a patient on the current study follows the Weibull distribution with shape parameter κ=1.22 and median survival time m_2. Let $\delta = (m_2/m_1)^\kappa$ be the hazard ratio, where m_1 is the median survival time of the historical control. Our aim is to test the following hypotheses:

$$H_0 : \delta \leq 1 \; vs \; H_0 : \delta > 1$$

with significance level of α=0.05 and power of 1-β=90% to detect an

				Type 1 error			Power		
Design	**κ**	**Test**	**At k^{th} interim look**	***k*=1**	***k*=2**	**total**	***k*=1**	***k*=2**	**total**
δ=1.5	0.5	*Z*(*t*)	Empirical	0.0067	0.0457	0.0524	0.3659	0.4293	0.7952
		S(*t*)	Empirical	0.0082	0.0435	0.0517	0.4104	0.3887	0.7992
		L(*t*)	Empirical	0.0066	0.046	0.0526	0.3545	0.4402	0.7947
			Nominal	0.0068	0.0435	0.0503	0.3686	0.4311	0.7997
	1	*Z*(*t*)	Empirical	0.0052	0.0484	0.0535	0.2874	0.5062	0.7936
		S(*t*)	Empirical	0.0066	0.0454	0.0521	0.3397	0.4594	0.799
		L(*t*)	Empirical	0.0058	0.0483	0.0541	0.2674	0.5245	0.7918
			Nominal	0.0055	0.045	0.0505	0.2912	0.5083	0.7995
	2.0	*Z*(*t*)	Empirical	0.0037	0.0503	0.054	0.185	0.6096	0.7946
		S(*t*)	Empirical	0.0054	0.0448	0.0502	0.2541	0.547	0.801
		L(*t*)	Empirical	0.0043	0.0498	0.0541	0.1655	0.6257	0.7912
			Nominal	0.0045	0.0463	0.0508	0.1997	0.5994	0.7991
δ=1.7	0.5	*Z*(*t*)	Empirical	0.0037	0.0457	0.0494	0.2611	0.556	0.8171
		S(*t*)	Empirical	0.0059	0.0455	0.0514	0.3054	0.4989	0.8043
		L(*t*)	Empirical	0.0041	0.0457	0.0497	0.2708	0.5467	0.8176
			Nominal	0.0054	0.0451	0.0505	0.2851	5144	0.7995
	1	*Z*(*t*)	Empirical	0.0023	0.047	0.0493	0.1612	0.6522	0.8134
		S(*t*)	Empirical	0.0052	0.0479	0.0531	0.2259	0.5806	0.8065
		L(*t*)	Empirical	0.0029	0.0466	0.0495	0.1753	0.5245	0.7918
			Nominal	0.0045	0.0462	0.0508	0.2049	0.5943	0.7992
	2	*Z*(*t*)	Empirical	0.0014	0.0486	0.0499	0.0607	0.7496	0.8104
		S(*t*)	Empirical	0.0046	0.0477	0.0523	0.1366	0.704	0.8031
		L(*t*)	Empirical	0.0024	0.0486	0.051	0.0858	0.7227	0.8085
			Nominal	0.0043	0.0471	0.0514	0.1259	0.6726	0.7985
δ=1.9	0.5	*Z*(*t*)	Empirical	0.0022	0.0452	0.0474	0.2041	0.629	0.833
		S(*t*)	Empirical	0.0051	0.0465	0.0516	0.2732	0.5402	0.8135
		L(*t*)	Empirical	0.003	0.0449	0.0479	0.2326	0.6015	0.8341
			Nominal	0.005	0.0455	0.0505	0.2574	0.542	0.7994
	1	*Z*(*t*)	Empirical	0.0011	0.0453	0.0464	0.0947	0.733	0.8276
		S(*t*)	Empirical	0.0046	0.0472	0.0518	0.1883	0.6239	0.8122
		L(*t*)	Empirical	0.003	0.0449	0.0479	0.2326	0.6015	0.8341
			Nominal	0.004	0.0465	0.0509	0.1803	0.6187	0.799
	2	*Z*(*t*)	Empirical	0.0002	0.047	0.0472	0.0067	0.8168	0.8235
		S(*t*)	Empirical	0.005	0.0484	0.0533	0.1005	0.704	0.8045
		L(*t*)	Empirical	0.0017	0.0468	0.0485	0.0513	0.771	0.8224
			Nominal	0.0044	0.0473	0.0516	0.1084	0.6899	0.7983

Table 2: Simulated empirical type I error and power of the two-stage SCPRT designs based on 100,000 simulation runs for sequential *Z*(*t*), log-rank *L*(*t*) and *S*(*t*) tests with a nominal type I error of 0.05 and power 80% (one-sided test).

At k^{th} interim look	***k*=1**	***k*=2**	***k*=3**	**total**
Type I error				
Empirical of *S*(*t*)	0.0028	0.0047	0.0422	0.0496
Nominal	0.0024	0.0046	0.0436	0.0506
Power				
Empirical of *S*(*t*)	0.171	0.2994	0.3886	0.8389
Nominal	0.1204	0.2533	0.4257	0.7994
Probability of stopping under null				
Empirical of *S*(*t*)	0.2574	0.3907	0.3519	1
Nominal	0.2626	0.3916	0.346	1
Probability of stopping under alternative				
Empirical of *S*(*t*)	0.1756	0.315	0.5094	1
Nominal	0.1315	0.28	0.5885	1

Table 3: Operating characteristics of the three-stage SCPRT design for test statistic *S*(*t*)based on 100,000 simulation runs under the Weibull distribution with uniform censoring distribution on $[t_f, t_a+t_f]$, and nominal type I error of 0.05 and power 80% for the example in Section 6.

alternative δ=1.714, which is calculated from by increasing 5 years median survival times of the historical control (m_1=9) to the current study (m_2=14). Given type I error α=.05, power of 90%, number of deaths of the historical control D_1=65, effect size δ=1.714, and the Weibull shape parameter κ=1.22, the number of events required for the current study for the *Z*(τ) test is calculated by

$$D_2(\tau)=\left\{\frac{[\log(1.714)]^2}{(1.645+1.282)^2}-65^{-1}\right\}^{-1}=54$$

The number of events required for the *S*(τ) test is calculated by

$$D_2(\tau)=\left\{\frac{9\times(1.714^{1/3}-1)^2}{(1.645+1.282)^2}-1.714^{2/3}\times 65^{-1}\right\}^{-1}=53.67,$$

which is 54 events too. Assume that the lengths of accrual and follow-up for the current study are t_a=5 and t_f=3, respectively, and the study duration is τ=8. Then the probability of having an event during the study for a subject on the current study can be calculated by numerical integration

$$p_2(\tau)=1\frac{1}{t_a}\int_{t_f}^{t_a+t_f}e^{-\log(2)\left(\frac{u}{m_2}\right)^{\kappa}}du=0.1985,$$

where κ=1.22 and m_2=14. Thus the number of patients required for the current study is $n_2 = D_2(\tau)/p_2(\tau) = 54/0.1985 = 273$. Suppose that the test statistic $S(t)$ will be used to monitor the trial, and 3 interim looks are planned at calendar times t_1=4, t_2=6 and t_3=8 years. Then the transformed information times can be calculated by

$$t_k^* = D(t_k)/D(\tau) = (1+R)I_k/(1+DI_k), \quad (14)$$

where $I_k = D_2(t_k)/D_2(\tau)$ and $R = \delta^{2/3}D_2(\tau)/D_1$, with $D_{1=65}$, $D_2(t_k) = n_2 p_2(t_k)$ and

$$p_2(t) = \frac{1}{t_a}\int_0^{t\wedge t_a}\{1 - S_2(t-u)\}du, \quad (15)$$

where $S_2(t) = e^{-log(2)\left(\frac{t}{m_2}\right)^{\kappa}}$. Thus, the information time calculated by equations (14) and (15) is t^*=(0.436, 0.773, 1), the lower and upper boundaries calculated by equation (12) are (a_1,a_2,a_3)=(-0.425, 0.307, 1.645) and (b_1,b_2,b_3)=(1.859, 2.236, 1.645), respectively, and the nominal critical p- values calculated by equation (13) are $(P_{a_1}, P_{a_2}, P_{a_3}) = (0.7398, 0.3634, 0.05)$ and $(P_{b_1}, P_{b_2}, P_{b_3}) = (0.0024, 0.0055, 0.05)$ for the lower and upper boundaries, respectively. To monitor the trial at k^{th} interim look, the survival data collected up to calendar time t_k from the current study combined with all data of the historical control to calculate the sequential test statistic $S(t_k)$ as described in Section 2, and the observed p-values

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

At k^{th} stage, we stop the trial for futility if $P_{S(t_k)} \geq P_{a_k}$, and stop the trial for efficacy if $P_{S(t_k)} \leq P_{b_k}$. The operating characteristics of the sequential test $S(t)$ for this example are given in Table 3.

Conclusion

We proposed two parametric sequential tests for group sequential trial de-sign against historical controls. Simulation results showed that the empirical power and type I error of the $S(t)$ test are close to those of the nominal levels, and it outperforms the $Z(t)$ test. Hence, we recommend using the $S(t)$ test for historical control trial designs under the Weibull model. We derived transformed information times t^*=(1+R)I/(1+RI) for both test statistics $Z(t)$ and $S(t)$. It is simple and convenient to use the transformed information time t^* to derive the sequential monitoring rule for the historical control trial design based on the SCPRT procedure. With this monitoring procedure, data from the current study are sequentially collected and com-pared with data from the historical control. This allows investigators to monitor the trial at any calendar time of enrollment or at a pre-specified number of events of an interim look. The number of events required for the current study can be calculated by a simple formula. Therefore, the study design is much simpler than that of the method for survival data proposed by Xiong et al. [12], in which information times of the sequential test statistic are random and depend on data instead of being predetermined. The maximum sample size of the sequential test is the same as that for the fixed sample test and the group sequential boundaries have analytical solutions. Therefore, the proposed group sequential procedure is effective and simple to use. For the study design purpose, we need the number of events from the historical control data only. However for the trial monitoring and final data analyses, we need full failure time data from the historical control study to calculate the sequential test statistic $Z(t_k)$ or $S(t_k)$. In practice, the historical control data are often available from previous trials done by the same institution or by the same sponsor. If there is no such historical control data available from the same institution, then we need to extract the relevant data from published literatures. Recently, Guyot et al. [20] have proposed a method to reconstructing the survival data from published Kaplan-Meier survival curves. Thus designing survival trials with historical controls are feasible by using control data from published literatures.

Finally, even though the sample size formula (5) and transformed information time (10) were derived for the $Z(t)$ test under the Weibull model, our simulation results showed that they also work well for the nonparametric log-rank test under the proportional hazard models. A rigorous derivation of these results for the log-rank test will be the future research.

Acknowledgment

This work was supported in part by the National Cancer Institute (NCI) support grant P30CA021765-35.

References

1. Xiong X (1995) A class of sequential conditional probability ratio tests. Journal of American Statistical Association 15: 1463-1473.
2. Pocock S (1976) Randomized versus historical controls: a compromise solution. PI Biom C 9: 245260.
3. Farewell VT, D'Angio GJ (1981) A simulated study of historical controls using real data. Biometrics 37: 169-176.
4. Gehan EA (1982) Design of controlled clinical trials: use of historical controls. Cancer Treat Rep 66: 1089-1093.
5. Fleming TR (1982) Historical controls, data banks, and randomized trials in clinical research: a review. Cancer Treat Rep 66: 1101-1105.
6. Green SB, Byar DP (1984) Using observational data from registries to compare treatments: the fallacy of omnimetrics. Stat Med 3: 361-373.
7. Makuch RW, Simon RM (1980) Sample size considerations for non-randomized comparative studies. J Chronic Dis 33: 175-181.
8. Dixon DO, Simon R (1988) Sample size considerations for studies comparing survival curves using historical controls. J Clin Epidemiol 41: 1209-1213.
9. Emrich LJ (1989) Required duration and power determinations for historically controlled studies of survival times. Stat Med 8: 153-160.
10. Korn EL, Freidlin B (2006) Conditional power calculations for clinical trials with historical controls. Stat Med 25: 2922-2931.
11. Chang MN, Shuster JJ, Kepner JL (1999) Group sequential designs for phase II trials with historical controls. Control Clin Trials 20: 353-364.
12. Xiong X, Tan M, Boyett J (2007) A sequential procedure for monitoring clinical trials against historical controls. Stat Med 26: 1497-1511.
13. Zhang S, Cao J, Ahn C (2010) Calculating sample size in trials using historical controls. Clin Trials 7: 343-353.
14. Cox DR, Oakes DV (1984) Analysis of Survival Data London: Chapman and Hall.
15. Tsiatis AA, Boucher H, Kim K (1995) Sequential methods for parametric survival models Biometrika 70: 165-173.
16. Sprott DA (1973) Normal likelihoods and relation to a large sample theory of estimation. Biometrika 60: 457-465.
17. Lawless JF (1982) Statistical methods for lifetime data. John Wiley and Sons, New York.
18. Collett D (2003) Modeling survival data in medical research, (2ndedn), London: Chapman and Hall.
19. Xiong X, Tan M, Boyett J (2003) Sequential conditional probability ratio tests for normalized test statistic on information time. Biometrics 59: 624-631.
20. Guyot P, Ades AE, Ouwens MJ, Welton NJ (2012) Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol 12: 9.