Adaptive design: estimation and inference with censored data in a semiparametric model

YU SHEN*

Department of Biostatistics, M. D. Anderson Cancer Center, Houston, TX 77030, USA
yshen@mdanderson.org

YI CHENG

Department of Mathematical Sciences, Indiana University at South Bend, IN 46634, USA

SUMMARY

In this article, we provide a method of estimation for the treatment effect in the adaptive design for censored survival data with or without adjusting for risk factors other than the treatment indicator. Within the semiparametric Cox proportional hazards model, we propose a bias-adjusted parameter estimator for the treatment coefficient and its asymptotic confidence interval at the end of the trial. The method for obtaining an asymptotic confidence interval and point estimator is based on a general distribution property of the final test statistic from the weighted linear rank statistics at the interims with or without considering the nuisance covariates. The computation of the estimates is straightforward. Extensive simulation studies show that the asymptotic confidence intervals have reasonable nominal probability of coverage, and the proposed point estimators are nearly unbiased with practical sample sizes.

Keywords: Confidence interval; Cox model; Martingale; Nuisance covariates; Point estimator; Self-designing.

1. INTRODUCTION

In a typical phase III trial, especially in studies of cancer and other chronic diseases, the primary goal is often to compare the duration of time to an event between patient groups on different treatment arms. The time to an event of interest may be censored due to loss to follow-up or the event not having occurred at the time of analysis. During the course of these long-term clinical trials, design adaptations sometimes are necessary on the basis of the observed data, without inflating the type I error rate while achieving the study goals. While attractive for its flexibility, the adaptive design makes it challenging to perform the final analyses, including the estimation of a treatment difference and its confidence interval at the end of trial. As expected, conventional analytic methods often lead to biased estimators due to the interim decisions within an adaptive design (Cheng and Shen, 2004).

Similarly, the estimation of the parameters is often difficult after the classical group sequential tests even with a fixed sample size. The classical group sequential designs were proposed to stop trials early for highly significant positive results in the fixed sample design framework (Pocock, 1977; O’Brien and
Fleming, 1979; Lan and DeMets, 1983). The basic motivations of the two types of group sequential designs are different. However, similar difficulties in estimation are found with either type of design due to sequential sampling. Following the classical sequential tests, Todd and Whitehead (1996) proposed a bias-adjusted MLE based on triangular test. Others, including Siegmund (1978), Tsiatis and others (1984), and Emerson and Fleming (1990), developed some orderings on the outcome space, either by the sample mean or by the analysis time at which the study terminates. These methods of estimation for the classical sequential designs with a fixed maximum sample size rely on a prespecified outcome space and, therefore, are not directly applicable to the adaptive design with a random maximum sample size.

Most adaptive designs and corresponding methods of analysis have been developed mainly for instantaneously observed continuous or binary outcomes; thus, the methods are not directly applicable to survival responses with staggered entry. In general, for time-to-event outcomes, the adaptive design and analysis procedures can be considerably complicated. With staggered entry and long-term follow-up, partial but increasing information becomes available from the study participants at successive monitoring times. The information is contained not only in the sample size but also in the number of events observed during the study. Within group sequential designs, Schäfer and Müller (2001) proposed a conditional rejection error probability function to make design modifications in an ongoing trial with censored survival data. More recently, Shen and Cai (2003) generalized the self-designing trial of Fisher (1998) and Shen and Fisher (1999) from immediately observed outcomes to censored survival outcomes with staggered entry for comparison of two treatment arms. However, the attention is still restricted to testing the equality of two survival distributions. As recognized by Schäfer and Müller (2001) and others, the problem of estimating the parameter of interest and its confidence interval at the end of the adaptive trials remains unsolved, especially with censored outcomes. Another limitation in the existing designs and inference is that the constructed test statistics do not consider other risk factors that are commonly available at baseline in clinical trials. When the effects of nuisance covariates are substantial, the sample size estimation based on the design without considering the nuisance covariates can be less efficient (Lagakos and Schoenfeld, 1984; Self and Mauritsen, 1988), and the estimator of the treatment effect can be biased (Gail and others, 1984; Struthers and Kalbfleisch, 1986).

Here, we describe the adaptive design and inference method within the framework of the Cox proportional hazards model, in the presence of covariates other than the treatment indicator in a randomized clinical trial. We construct the bias-adjusted estimator of the hazard ratio between the treatment arms and its confidence interval at the end of the trial, with or without other covariates adjustment in the model. The final statistic for testing the treatment effect can be derived as a weighted average of linear rank statistics, while adjusting for the effects of nuisance covariates via the semiparametric Cox proportional hazards function. In Section 2, we present the estimation and inference procedures for the adaptive design with and without nuisance covariates. We summarize the simulation studies in Section 3, and illustrate an example in Section 4. We make general remarks in Section 5.

2. Inference and estimation of treatment effect

2.1 Notation and setting

We consider the setting where patient entry is staggered, random, and independent of the survival and censoring times. After an initial accrual period of $t_a$, patients are followed for an additional period of time, $t_f$. Assume that the two treatment groups under study are entered sequentially and allocated symmetrically. The total number of interim analyses $m$ does not have to be fixed in advance. Let $A_i$ denote the time of entry, $C_i$ denote the time from entry to censoring, $V_i$ denote the time from entry to failure, $Z_i$ be the indicator variable for the treatment group, and $W_i$ be a vector of nuisance variables for the $i$th patient, $i = 1, 2, \ldots$. A key assumption for the method is that the time of entry into the study is statistically
independent of the time to failure (or censoring), and the random variables $A$, $V$, and $C$ are conditionally independent given covariates (Tsiatis, 1981; Tsiatis and others, 1985). Data collected on the $i$th individual are described by independent and identically distributed vectors, $\{(A_i, V_i, C_i, Z_i, W_i), i = 1, \ldots\}$.

When data are reviewed at time $t$, we observe the time to failure or censoring, $X_i(t)$, and the failure indicator $\Delta_i(t)$ for the $i$th individual, where $X_i(t) = \max\{\min(V_i, t - A_i, C_i), 0\}$, and $\Delta_i(t) = 1$ if $V_i \leq \min(t - A_i, C_i)$ and 0 otherwise. The following variables can be represented as identically and independently distributed random vectors $\{X_i(t), \Delta_i(t), Z_i, W_i\}$ for $i = 1, 2, \ldots$. Let $\mathcal{F}_t = \sigma\{X_i(u), \Delta_i(u), Z_i, W_i, 0 < u < t, i = 1, \ldots\}$ define a filtration at time $t$. Thus, $\mathcal{F}_{t_j}$ is the $\sigma$-field generated by the cumulated data for each block up to the $j$th look, where $j = 1, 2, \ldots, m$.

### 2.2 Sequential inference for treatment effect only

First, we consider the simple two-arm randomized design without nuisance covariates. To measure the treatment effect, we assume the semiparametric proportional hazards model (Cox, 1972)

$$
\lambda_1(t|Z = z) = e^{\theta z} \lambda_0(t),
$$

(2.1)

where $Z = 1$ for a patient in the intervention arm, $Z = 0$ in the control, and $\theta$ is the log-relative risk of the event. The one-sided hypotheses to be tested is as follows—$H_0: \theta = \theta_0$ against the alternative $H_1: \theta > \theta_0$ with type I error rate $\alpha$ and power $1 - \beta$ at $\theta = \delta$, where $\delta > 0$. Without loss of generality, let $\theta_0 = 0$.

Shen and Cai (2003) introduced a weighted sum of rank test statistics that can update the sample size based on the observed survival data with staggered entry. The asymptotic distribution had been derived for the proposed test statistic under sequential monitoring in this setting. The trial is assessed periodically based on the observed survival data with staggered entry. The asymptotic independent increments property of independent given covariates (Tsiatis, 1981; Tsiatis and others, 1985). Data collected on the $i$th individual are described by independent and identically distributed vectors, $\{(A_i, V_i, C_i, Z_i, W_i), i = 1, \ldots\}$.

Sequential inference for treatment effect only

Shen and Cai (2003) introduced a weighted sum of rank test statistics that can update the sample size based on the observed survival data with staggered entry. The asymptotic distribution had been derived for the proposed test statistic under sequential monitoring in this setting. The trial is assessed periodically based on the observed survival data with staggered entry. The asymptotic independent increments property of independent given covariates (Tsiatis, 1981; Tsiatis and others, 1985). Data collected on the $i$th individual are described by independent and identically distributed vectors, $\{(A_i, V_i, C_i, Z_i, W_i), i = 1, \ldots\}$.

### 2.2 Sequential inference for treatment effect only

First, we consider the simple two-arm randomized design without nuisance covariates. To measure the treatment effect, we assume the semiparametric proportional hazards model (Cox, 1972)

$$
\lambda_1(t|Z = z) = e^{\theta z} \lambda_0(t),
$$

(2.1)

where $Z = 1$ for a patient in the intervention arm, $Z = 0$ in the control, and $\theta$ is the log-relative risk of the event. The one-sided hypotheses to be tested is as follows—$H_0: \theta = \theta_0$ against the alternative $H_1: \theta > \theta_0$ with type I error rate $\alpha$ and power $1 - \beta$ at $\theta = \delta$, where $\delta > 0$. Without loss of generality, let $\theta_0 = 0$.

Shen and Cai (2003) introduced a weighted sum of rank test statistics that can update the sample size based on the observed survival data with staggered entry. The asymptotic distribution had been derived for the proposed test statistic under sequential monitoring in this setting. The trial is assessed periodically after observing every $B_j$ events between $t_{j-1}$ and $t_j$, where $(t_1 < t_2 < \cdots < t_m)$ are determined as the time when one observes $B_j$ events since $t_{j-1}$, and $B_j$ is often a prefixed constant integer. At the $j$th interim analysis at time $t_j$, $n_j$ individuals have entered into the study, and $B_j$ events have been observed from $t_{j-1}$; a linear rank test statistic based on all cumulated data up to time $t_j$, $U(t_j, \theta_0)$, is computed under the null hypothesis. Specifically, the statistic at calendar time $t_j$ can be written as

$$
U(t_j, \theta_0) = \sum_{i=1}^{n_j} \int_0^{t_j} \{Z_i(u) - \bar{Z}(t_j, u)\} N_i(t_j, du),
$$

where $N_i(t, u) = I(X_i(t) \leq u, \Delta_i(t) = 1)$, $\bar{Z}(t_j, u) = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_i(t, u)Z_i^\theta e^{\theta Z_i}(r = 0, 1)$, and $Y_i(t, u) = I\{X_i(t) \geq u\}$.

It has been proven that the increments $U(t_1, \theta_0), U(t_2, \theta_0) - U(t_1, \theta_0), \ldots, U(t_j, \theta_0) - U(t_{j-1}, \theta_0)$ are asymptotically independent for prefixed times (Tsiatis, 1981; Slud, 1984, and others). Gail and others (1982) found that the asymptotic joint distribution of the linear rank statistics without nuisance covariates, evaluated at times defined by observing prespecified number of events, has the same structure of independent increments as prefixed times. Tsiatis and others (1985) further confirmed this independent increment structure under the Cox proportional hazards model with nuisance covariates. A theoretical justification using stopping times can be found in Appendix A for the asymptotic independent increments property of $U(t_j, \cdot)$ for $j = 1, \ldots, m$.

Under the null hypothesis, the distribution of $n_j^{1/2}\{U(t_j, \theta_0) - U(t_{j-1}, \theta_0)\}$ converges to a normal distribution with mean 0 and variance $\sigma_j^2 - \sigma_{j-1}^2$, where $\sigma_j^2$ is the variance of $n_j^{1/2}U(t_j, \theta_0)$. Note that $\sigma_j^2$ can be estimated using the formula of Fleming and Harrington (1991, p 261). Thus, under $H_0$ the
standardized information cumulated during \((t_{j-1}, t_j)\) can be expressed as

\[
S(t_j) = n_j \frac{1}{2} \left( U(t_j, \theta_0) - U(t_{j-1}, \theta_0) \right) / \left( \hat{\sigma}_j^2 - \hat{\sigma}_{j-1}^2 \right)^{1/2}, \quad j = 1, 2, \ldots
\]

which is asymptotically normal, with mean 0 and \(\text{Var}(S(t_j)) = 1\) by Slutsky’s theorem.

Recall that the final test statistic \(T_m\) in a self-designing trial is constructed based on the weighted standardized \(S(t_j)\) as

\[
T_m = \sum_{j=1}^{m} w_j S(t_j),
\]

where \(w_j = w_j(S(t_1), \ldots, S(t_{j-1}))\) is a nonnegative weight function consisting of data up to the \(j\)th step, and \(\sum_{j=1}^{m} w_j^2 = 1\) (Shen and Cai, 2003). Under the null hypothesis, \(T_m\) is asymptotically normally distributed with mean 0 and variance 1.

In a self-designing clinical trial, it is important to assess futility at the interims for ethical and economic reasons. For this purpose, we use the Wald-type constant likelihood futility boundary proposed by Shen and Cai (2003) at time \(t_j\) defined by

\[
f(t_j) = \sqrt{n_j} \sigma_j \left\{ \sqrt{n_j} \delta \sigma_j / 2 + \log \left( \frac{\beta}{1 - \alpha} \right) / \sqrt{n_j} \delta \sigma_j \right\}.
\]

If the linear rank test statistic at \(t_j\), \(U(t_j, \theta_0) \leq f(t_j)\), the trial is terminated and the null hypothesis is accepted, because of concerns that the intervention may be ineffective or inferior to the standard one. Given the sequential procedure, we can explicitly express the probability of rejecting the null hypothesis at stage \(m\) as a probability of not terminating the trial for futility at any of the first \((m - 1)\) interim analyses, and the final test statistic being greater than the critical value \(z_\alpha\), which follows

\[
P_1 = P(U(t_1, \theta_0) > f(t_1), \ldots, U(t_{m-1}, \theta_0) > f(t_{m-1}), T_m > z_\alpha) \leq P(T_m > z_\alpha | \theta = \theta_0) = \alpha,
\]

under the null hypothesis. Thus, the probability of rejecting the null hypothesis under \(H_0\) is below the nominal level, \(\alpha\).

If the trial is not terminated by futility, i.e., \(U(t_j, \theta_0) > f(t_j)\), the recruitment is continued, and a weight function is established based on the cumulated data prior to the \(j\)th step, as long as \(\sum_{i=1}^{j} w_i^2 < 1\). The procedure is iterated until step \(m\), if \(\sum_{j=1}^{m-1} w_j^2 < 1\) and \(\sum_{j=1}^{m} w_j^2 \geq 1\). In other words, all the weight function is spent at step \(m\), and the trial is terminated at the calendar time \(t_m\). Specifically, the weight function is constructed as an inverse function of the additional number of events required to achieve the specified power \((1 - \beta)\) at step \(j (j > 1)\), denoted by \(d(t_j^*) - d(t_{j-1})\); thus,

\[
w_j = \sqrt{\frac{B_j (1 - \sum_{i=1}^{j-1} w_i^2)}{d(t_j^*) - d(t_{j-1})}}, \quad \text{for } j > 1,
\]

where \(t_j^*\) is the calendar time to exit the trial. Note that \(t_j^*\) is different from \(t_j\) in general. By solving the conditional power equation, we obtain

\[
d(t_j^*) - d(t_{j-1}) = \left\{ \frac{z_\alpha - \sum_{i=1}^{j-1} w_i S_i}{\sqrt{1 - \sum_{i=1}^{j-1} w_i^2}} + z_\beta \right\}^2 / \hat{\sigma}_{j-1}^2.
\]
where $z_\alpha$ and $z_\beta$ denote the upper $\alpha$- and $\beta$-standard normal percentiles, $\hat{\sigma}_j^2$ is the estimated variance of the treatment indicator $Z$, and $\hat{\theta}_{j-1}$ is the estimated hazard ratio based on the partial likelihood using data up to the $(j-1)$th block. The weight at the first step is often assigned a fixed fraction based on the design parameters (Shen and Fisher, 1999; Cheng and Shen, 2004).

With the fixed sample design, the maximum partial likelihood estimate $\hat{\theta}$ solved from the score equation, $U(\theta, t_m) = 0$, at the end of the trial is a consistent estimate of $\theta$. However, with the adaptive sequential designs, this estimator can be biased due to interim stopping rules. To derive an asymptotically unbiased estimator and its confidence interval for $\theta$, we need to construct a "pivotal" quantity of the unknown parameter $\theta$ and the final test statistic, so that the asymptotic distribution of the "pivot" is known both under the null and under the alternative hypotheses (Cheng and Shen, 2004).

One difficulty involving censored survival data is that the final test statistic is proposed through the score statistic at each stage. Applying the martingale central limit theorem, the cumulated score statistic evaluated at time $t_j$, $n_j^{-1/2}U(t_j, \theta)$, asymptotically behaves like a Gaussian process with mean value $n_j\theta\hat{\sigma}_j^2$ and variance $\hat{\sigma}_j^2$. The tightness of the process $n_j^{-1/2}U(t_j, \theta)$ has been established for the Cox model with covariates by Bilias and others (1997). Under the alternatives (for any $\theta > 0$), the normalized process of $S(t_j)$, $S(t_j) - \theta \sqrt{n_j(\sigma_j^2 - \hat{\sigma}_j^2)}$, converges weakly to a zero-mean Guassian process. We, therefore, define the pivot as

$$V_m = \sum_{j=1}^m w_j \left( S(t_j) - \theta \sqrt{n_j(\hat{\sigma}_j^2 - \hat{\sigma}_{j-1}^2)} \right).$$

Through a proof similar to that of Shen and Cai (2003), we can verify that $V_m$ asymptotically converges to a standard normal distribution for any $\theta$ (under both the null and alternative hypotheses), using the properties of the martingale theorem and the characteristic function. Based on the above distribution property for $V_m$, a two-sided asymptotic confidence interval for $\theta$ at the end of the trial can be constructed at significance level $100(1 - 2\alpha)$ from

$$\Pr \left( -z_{1-\alpha} \leq T_m - \sum_{j=1}^m w_j \theta \sqrt{n_j(\hat{\sigma}_j^2 - \hat{\sigma}_{j-1}^2)} \leq z_{1-\alpha} \right) \approx 1 - 2\alpha,$$

and leads to the form

$$\left( \frac{T_m - z_{1-\alpha}}{\sum_{j=1}^m w_j \sqrt{n_j(\hat{\sigma}_j^2 - \hat{\sigma}_{j-1}^2)}}, \frac{T_m + z_{1-\alpha}}{\sum_{j=1}^m w_j \sqrt{n_j(\hat{\sigma}_j^2 - \hat{\sigma}_{j-1}^2)}} \right).$$

The distribution of $V_m$ also yields to a bias-adjusted moment estimator for $\theta$, which is also the center of the constructed confidence interval,

$$\hat{\theta} = \frac{T_m}{\sum_{j=1}^m w_j \sqrt{n_j(\hat{\sigma}_j^2 - \hat{\sigma}_{j-1}^2)}}.$$

Using the arguments similar to those in Cheng and Shen (2004), we can prove the consistency of $\hat{\theta}$ in terms of block size. The practical implication of this property is that the block size for the adaptive design should be moderate to have a consistent estimator.

To ensure the martingale property of $V_m$, one more block of data (or at least one failure) should be observed to spend all the left weight at the $m$th step ($m = j + 1$), when a trial is terminated at the $j$th step due to futility, or when the conditional power at the $j$th step is greater than $1 - \beta$. If we allow the
Adaptive design

311

trial to stop at the \( j \)th step and accept \( H_0 \) due to futility, the event \( \{m = j\} \) may no longer be \( \mathcal{F}_{m-1} \) measurable. Therefore, \( V_m \) with \( m = j \) does not have an asymptotic standard Gaussian distribution. As the associated editor suggested, it is possible to use empirical process theory instead of martingale theory to derive the asymptotic distribution of \( V_m \) using data observed up to step \( j \). However, the variance of \( \hat{\theta} \) may be increased slightly without the extra step. For the futility termination, the decision to accept \( H_0 \) should remain the same even with the extra block of observed data to maintain the logical consistency, as discussed by Cheng and Shen (2004).

2.3 Sequential inference in the presence of nuisance covariates

It has been widely recognized that the use of important risk factors can increase the power of the inference in the design and reduce bias in the estimation of the treatment effect in the analysis of clinical trials. Therefore, an important generalization of the adaptive design for censored survival outcomes is directed to the comparison of the survival distributions with adjustment for other covariates. The Cox regression model (Cox, 1972) is a most commonly used semiparametric model for adjusting nuisance covariates, while the primary interest is to assess the treatment effect.

Let \( \gamma \) define the covariate coefficient vector for the nuisance covariate vector, \( W \). Under the proportional hazards model assumption, the hazard function, given all available covariates, is as follows:

\[
\hat{\lambda}(t|Z, W) = \hat{\lambda}_0(t) \exp(\theta Z) \exp(\gamma^T W).
\]

With staggered entry, the asymptotic theory has been developed for the joint distribution of a sequence of score statistics evaluated at different times under the Cox model for the classical group sequential designs (Jennison and Turnbull, 1997; Gu and Ying, 1995; Tsiatis and others, 1985). Specifically, the sequence of score statistics has an asymptotical joint multinormal distribution and an independent increment property. At the \( j \)th interim analysis at time \( t_j \), the score process for testing the null hypothesis, \( \theta = \theta_0 \), is denoted by \( U(t_j, \theta_0, \hat{\gamma}_j) \), where

\[
U(t_j, \theta_0, \gamma) = \sum_{i=1}^{n_j} \int_0^{t_j} \left( Z_i(u) - \sum_k Z_k(u) \exp(\gamma^T W_k(u)) I\{X_k(t_j) \geq u\} \right) N_i(t_j, du),
\]

and \( \hat{\gamma}_j \) is solved from the constrained partial likelihood equation with \( \theta = \theta_0 \) based on data observed up to the review time \( t_j \). That is,

\[
H(t_j, \theta_0, \gamma) = \sum_{i=1}^{n_j} \int_0^{t_j} \left( W_i(u) - \sum_k W_k(u) \exp(\gamma^T W_k(u)) I\{X_k(t_j) \geq u\} \right) N_i(t_j, du) = 0.
\]

Denote \( \gamma_0 \) to be the true value of \( \gamma \). Tsiatis and others (1985) showed that the finite-dimensional distributions of \( n^{-1/2}U(t_j, \theta_0, \hat{\gamma}_j) \) and \( n^{-1/2}U(t_j, \theta_0, \gamma_0) \) are asymptotically equivalent and the limiting process is a time-rescaled Brownian motion, if \( Z_i \) and \( W_i \) are independent. The results have been extended by Gu and Ying (1995) to more general cases with dependent \( Z \) and \( W \). Based on these established properties, \( U(t, \theta_0, \hat{\gamma}_j) \) has independent increments with any type of joint distribution for \( Z \) and \( W \). At two different review times \( t_j \) and \( t_k \), \( n^{-1/2}(U(t_j, \theta_0, \hat{\gamma}_j), U(t_k, \theta_0, \hat{\gamma}_k)) \) converge in distribution to a bivariate Gaussian
process with mean 0 and covariance function $\sigma^2_{\zeta}(t_j)$ for $t_j < t_k$, and

$$
\sigma^2_{\zeta}(t_j) = I_{zz}(t_j) - I_{wu}(t_j)I_{ww}(t_j)^{-1}I^T_{wu}(t_j),
$$

where

$$
I_{zz}(t) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{t} (Z_i(u) - \tilde{Z}(t, u))^2 \exp(\gamma_0^T W_i(u))Y_i(t, u)d\Lambda_0(u),
$$

$$
I_{wu}(t) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{t} (W_i(u) - \tilde{W}(t, u))^2 \exp(\gamma_0^T W_i(u))Y_i(t, u)d\Lambda_0(u),
$$

$$
I_{ww}(t) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{t} (Z_i(u) - \tilde{Z}(t, u)) \times (W_i(u) - \tilde{W}(t, u))^T \exp(\gamma_0^T W_i(u))Y_i(t, u)d\Lambda_0(u),
$$

where $\tilde{Z}(t, u)$ and $Y_i(t, u)$ are defined in Section 2.1, and

$$
\tilde{W}(t, u) = \frac{\sum_{i=1}^{n} W_i(u) \exp(\gamma_0^T W_i(u))Y_i(t, u)}{\sum_{i=1}^{n} \exp(\gamma_0^T W_i(u))Y_i(t, u)}.
$$

The consistent estimator of $\sigma^2_{\zeta}(t_j)$ can be obtained by substituting $\gamma_0$ and $\Lambda_0(u)$ in the above formulae by $\hat{\gamma}_j$ and

$$
\hat{\Lambda}_0(u; t_j) = \sum_{i=1}^{n_j} \frac{\Delta_i(t_j)I\{X_i(t_j) \leq u\}}{\sum_{i=1}^{n_j} \exp(\hat{\gamma}_j^T W_i(u))I\{X_i(t_j) \geq X_i(t_j)\}}.
$$

The proof for the consistency of $\hat{\sigma^2_{\zeta}}(t_j)$ was outlined in Gu and Ying (1995).

The construction of the final test statistic in this setting then becomes straightforward. We can follow a procedure similar to that described in Section 2.2 without nuisance covariates,

$$
T_m(W) = \sum_{j=1}^{m} \tilde{\omega}_j S(t_j, \hat{\gamma}_j),
$$

where

$$
S(t_j, \hat{\gamma}_j) = n_j^{-1} \{U(t_j, \theta_0, \hat{\gamma}_j) - U(t_j-1, \theta_0, \hat{\gamma}_{j-1})\}/(\hat{\sigma^2_{\zeta}}(t_j) - \hat{\sigma^2_{\zeta}}(t_{j-1}))^{1/2}
$$

and $\tilde{\omega}_j$ depends on all the information, including nuisance covariate data observed up to $t_{j-1}$. We will outline the derivation of $\tilde{\omega}_j$ next. Note that the constructed test statistic, $T_m(W)$, corresponds to testing the null hypothesis for $\theta = \theta_0$, while the variance estimator, $\hat{\sigma^2_{\zeta}}(t_j)$, contains information from both the nuisance covariates and the treatment indicator.

The weight function, $\tilde{\omega}_j$, for $j > 1$ is estimated iteratively using observed data prior to the $j$th look. For $j = 1$, as usual, we often assign a fixed fraction based on the design parameters. To ensure the specified power $1 - \beta$ given the accumulated data up to step $(j - 1)$, we will solve the additional number of events needed, $d(t_j^*) - d(t_{j-1})$, from the conditional power equation,

$$
P \left( \sum_{i=1}^{j-1} \tilde{\omega}_i S(t_i, \hat{\gamma}_i) + \tilde{\omega}_j S(t_j^*, \gamma) > z_\alpha|\mathcal{F}_{j-1} \right) = 1 - \beta,
$$

where $S(t_j^*, \gamma)$ is the standardized increment of the covariate-adjusted score statistic between $t_{j-1}$ and $t_j^*$. We replace $\theta$ and $\gamma$ by the estimated $\hat{\theta}_{j-1}$ and $\hat{\gamma}_{j-1}$, respectively, using the observed data up to step
(j - 1) on the right side of the following inequality:

\[
P\left( S(t_j^*, \gamma) - \hat{\theta}_{j-1} \sqrt{n_j(\sigma_z^2(t_j) - \sigma_z^2(t_{j-1}))} \right) > \frac{z_\alpha - \sum_{i=1}^{j-1} \tilde{w}_i S(t_i, \hat{\gamma}_i)}{\tilde{w}_j} - \hat{\theta}_{j-1} \sqrt{n_j(\sigma_z^2(t_j) - \sigma_z^2(t_{j-1})) | F_{j-1}} = 1 - \beta,
\]

where \(\sqrt{n_j(\sigma_z^2(t_j) - \sigma_z^2(t_{j-1}))}\) can be approximated by \(\pi_{z_\alpha} \sqrt{d(t_j^*) - d(t_{j-1})}\), and \(\pi_z^2\) is the variance of treatment indicator \(Z\) (Tsiatis and others, 1985). This approximation is based on a condition that the treatment assignment is independent of nuisance covariates \(W\). In randomized controlled clinical trials, this condition is easily satisfied. Furthermore, this approximation is used to construct the weight function only. The consistent estimator of \(\sigma_z^2(t_j)\) used in the final test statistic is general enough to handle a possible dependence between \(Z\) and \(W\).

Assuming that the \(j\)th step is the last step of the trial, then \(\tilde{w}_j = \sqrt{1 - \sum_{i=1}^{j-1} \tilde{w}_i^2}\). By solving the above conditional power equation, we obtain \(d(t_j^*) - d(t_{j-1})\) as a function depending on data observed prior to the \(j\)th step

\[
d(t_j^*) - d(t_{j-1}) = \left\{ \frac{z_\alpha - \sum_{i=1}^{j-1} \tilde{w}_i S(t_i, \hat{\gamma}_i)}{\sqrt{1 - \sum_{i=1}^{j-1} \tilde{w}_i^2}} + z_\beta \right\}^2 \tilde{\sigma}_z^2 \hat{\sigma}_z^2.
\]

With an estimated additional number of events required between \(t_{j-1}\) and \(t_j^*\), the weight function can be constructed exactly the same as in (A.1).

Using arguments similar to those used for cases without nuisance covariates, we can prove the asymptotic properties of \(T_m(W)\). (More details are provided in Appendix B.) Specifically, the pivotal function,

\[
V_m(W) = \sum_{j=1}^m \tilde{w}_j \left( S(t_j, \hat{\gamma}_j) - \hat{\theta} \sqrt{n_j(\hat{\sigma}_z^2(t_j) - \hat{\sigma}_z^2(t_{j-1}))} \right),
\]

follows an asymptotic standard normal distribution for any \(\theta\). Thus, the bias-adjusted moment estimate for \(\theta\) and its corresponding 100\((1 - \alpha)\)% confidence interval can be derived as

\[
\hat{\theta} = \frac{T_m(W)}{\sum_{j=1}^m \tilde{w}_j \sqrt{n_j(\hat{\sigma}_z^2(t_j) - \hat{\sigma}_z^2(t_{j-1}))}}, \quad \text{and} \quad T_m(W) = \frac{T_m(W) - z_{1-\alpha}}{\sum_{j=1}^m \tilde{w}_j \sqrt{n_j(\hat{\sigma}_z^2(t_j) - \hat{\sigma}_z^2(t_{j-1}))}} \quad \text{and} \quad T_m(W) + z_{1-\alpha} \sum_{j=1}^m \tilde{w}_j \sqrt{n_j(\hat{\sigma}_z^2(t_j) - \hat{\sigma}_z^2(t_{j-1}))}.
\]

Under the null hypothesis, \(\theta = 0\), \(V_m(W)\) reduces to the test statistic, \(T_m(W)\), which follows an asymptotic standard normal distribution. The futility stopping rule is derived the same way as in Section 2.2, replacing \(\sigma_j\) by \(\sigma_z(t_j)\) in (2.3).

3. Simulation study

We conduct a series of simulation studies to evaluate the performance of the proposed adaptive design inferences, including the test and estimation under either the null or the alternative hypotheses. For different
underlying survival distributions, we compare the power/size properties of the test with or without a nuisance covariate adjustment using the proportional hazards model in the adaptive designs, and compare the properties with the corresponding standard test statistic with or without a nuisance covariate in the fixed sample designs. We also obtain the naive estimates for both the treatment effect and the nuisance covariate in the adaptive designs, the adjusted estimates for the treatment effect with and without the nuisance covariate at the end of the adaptive trial, and compare them with the corresponding naive estimates in the fixed sample design, which should be asymptotically unbiased estimates under the correctly specified models.

The empirical type I error rates and power estimates are based on simulating 1000 independent clinical trials under different designs. All tests performed are one-sided. We simulate patient survival times from the proportional hazards model with one treatment indicator variable, $Z$, and one nuisance covariate, $W$. The two covariates are dichotomous, independent, and follow a Bernoulli distribution with probability of 0.5. The relative risk of the covariate $Z$ is $\theta$, and the relative risk of $W$ is $\gamma$. Assume that patients enter the trial with a Poisson process at a fixed rate per unit time. The baseline survival follows an exponential distribution.

As investigated by Shen and Cai (2003), different approaches to adjusting the number of events in a clinical trial include varying the accrual duration, accrual rate, or follow-up duration after accrual. Due to limited space, we consider only the first type of adjustment in the simulations. Specifically, this adaptive design allows for the accrual duration to be changed from 2 to 4 years, whereas the fixed sample design is based on an accrual duration of 2 years and an additional follow-up period of 2 years after accrual. Assume a fixed accrual rate of 126 patients per year. This setting requires 88 events (252 patients) to achieve 90% power with a type I error rate of 0.025 to detect an expected log-relative risk of $\delta = \log(2) = 0.693$ (baseline hazard rate of 0.2 in the control arm). With the fixed hazard rate in the control group, we assume various hazard rates in the treatment group, which leads to the log-hazard ratio, $\theta$, varying from 0.600 to 0.693. We consider the effect of the nuisance covariate, described by the log-hazard ratio, $\gamma$, taking values from 0, 0.5, and 0.8 to 1.2.

The number of events in the first interim analysis are specified as 40% of the number of events required in the fixed sample design (denoted as $B_1$). For the subsequent analyses, we use a constant block size of events of 5 or 10. In each simulation, a usual logrank statistic without adjusting the nuisance covariate and a statistic adjusting for the covariate in the Cox model for testing the treatment effect are computed based on the fixed sample design at the prespecified termination time, and compared with the adaptive test statistics with and without adjusting the covariate in terms of the type I error rate, power, and average number of events. We evaluate the performance of the proposed point estimate (denoted by “Adj. Est, $\theta$” in the tables) and the corresponding coverage probability of the 95% confidence interval (denoted by “Adj. 95% CP, $\theta$” in the tables) with or without adjusting the nuisance covariate, and compare the estimate with the corresponding naive estimate of $\theta$, which is obtained by simply solving the score equation at the end of the trial without considering the interim analyses (denoted by “Naive Est, $\theta$” in the tables).

Table 1 summarizes the empirical type I error rates and the performance of the estimates for $\theta$ and $\gamma$ under the null hypothesis. The results show that the proposed adaptive statistics are quite conservative compared to the standard tests with or without adjusting the nuisance covariate in terms of type I error rate. As expected, the naive estimators of $\theta$ obtained from the score equations at the end of the adaptive trials are somewhat biased compared to the adjusted estimators. The coverage probability of 95% confidence intervals given in Section 2 is fairly accurate across all cases with or without adjusting for the nuisance covariate. The empirical results are consistent with the observations of Struthers and Kalbfleisch (1986) that the estimators of $\theta$ under both (2.1) and (2.5) are correct when $\theta = 0$. In other words, if $\theta = 0$, omitting the nuisance covariate, $W$, has little impact on the inference for the treatment effect when $W$ and $Z$ are independent. Therefore, the corresponding bias-adjusted estimators and their confidence intervals for the adaptive designs are comparable with or without adjusting for the nuisance covariate. At the end of
Table 1. Empirical size and average number of events (ANE), bias-adjusted estimator and its 95% confidence interval for a design with one-sided $\alpha = 0.025$, $1 - \beta = 0.9$, and alternative hypothesis $H_1: \delta = 0.693$ with a baseline hazard of 0.2

<table>
<thead>
<tr>
<th></th>
<th>Logrank $B = 5$</th>
<th>Logrank $B = 10$</th>
<th>Cox PH model $B = 5$</th>
<th>Cox PH model $B = 10$</th>
<th>Fixed sample design Logrank</th>
<th>Fixed sample design Cox model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma = 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical size</td>
<td>0.014</td>
<td>0.021</td>
<td>0.014</td>
<td>0.019</td>
<td>0.033</td>
<td>0.032</td>
</tr>
<tr>
<td>ANE</td>
<td>79.36</td>
<td>87.72</td>
<td>79.41</td>
<td>88.05</td>
<td>112.73</td>
<td>112.73</td>
</tr>
<tr>
<td>Adj. Est, $\theta$</td>
<td>0.011</td>
<td>0.014</td>
<td>0.011</td>
<td>0.014</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% CP, $\theta$</td>
<td>0.944</td>
<td>0.940</td>
<td>0.941</td>
<td>0.940</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naive Est, $\theta$</td>
<td>-0.087</td>
<td>-0.070</td>
<td>-0.087</td>
<td>-0.070</td>
<td>-0.006</td>
<td>-0.006</td>
</tr>
<tr>
<td>Naive Est, $\gamma$</td>
<td>0.019</td>
<td>0.014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma = 0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical size</td>
<td>0.022</td>
<td>0.022</td>
<td>0.023</td>
<td>0.019</td>
<td>0.023</td>
<td>0.023</td>
</tr>
<tr>
<td>ANE</td>
<td>77.62</td>
<td>85.64</td>
<td>76.72</td>
<td>84.93</td>
<td>94.61</td>
<td>94.61</td>
</tr>
<tr>
<td>Adj. Est, $\theta$</td>
<td>0.010</td>
<td>0.013</td>
<td>0.011</td>
<td>0.014</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% CP, $\theta$</td>
<td>0.922</td>
<td>0.924</td>
<td>0.916</td>
<td>0.923</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naive Est, $\theta$</td>
<td>-0.099</td>
<td>-0.087</td>
<td>-0.100</td>
<td>-0.088</td>
<td>0.007</td>
<td>0.009</td>
</tr>
<tr>
<td>Naive Est, $\gamma$</td>
<td>0.512</td>
<td>0.504</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma = 0.8$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical size</td>
<td>0.016</td>
<td>0.019</td>
<td>0.017</td>
<td>0.023</td>
<td>0.028</td>
<td>0.034</td>
</tr>
<tr>
<td>ANE</td>
<td>76.76</td>
<td>83.97</td>
<td>75.46</td>
<td>83.07</td>
<td>86.17</td>
<td>86.17</td>
</tr>
<tr>
<td>Adj. Est, $\theta$</td>
<td>0.012</td>
<td>0.016</td>
<td>0.016</td>
<td>0.020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% CP, $\theta$</td>
<td>0.934</td>
<td>0.932</td>
<td>0.934</td>
<td>0.933</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naive Est, $\theta$</td>
<td>-0.097</td>
<td>-0.080</td>
<td>-0.091</td>
<td>-0.074</td>
<td>0.007</td>
<td>0.004</td>
</tr>
<tr>
<td>Naive Est, $\gamma$</td>
<td>0.827</td>
<td>0.823</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma = 1.2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical size</td>
<td>0.014</td>
<td>0.019</td>
<td>0.022</td>
<td>0.021</td>
<td>0.028</td>
<td>0.033</td>
</tr>
<tr>
<td>ANE</td>
<td>76.37</td>
<td>83.97</td>
<td>75.51</td>
<td>82.75</td>
<td>77.45</td>
<td>77.45</td>
</tr>
<tr>
<td>Adj. Est, $\theta$</td>
<td>0.013</td>
<td>0.020</td>
<td>0.016</td>
<td>0.026</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% CP, $\theta$</td>
<td>0.936</td>
<td>0.938</td>
<td>0.929</td>
<td>0.936</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naive Est, $\theta$</td>
<td>-0.092</td>
<td>-0.075</td>
<td>-0.092</td>
<td>-0.071</td>
<td>0.007</td>
<td>0.005</td>
</tr>
<tr>
<td>Naive Est, $\gamma$</td>
<td>1.221</td>
<td>1.218</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

PH, proportional hazards; Adj. Est, adjusted estimator; CP, coverage probability.

the adaptive design with the adjustment for the nuisance covariate, we also calculate the naive estimates for $\gamma$, which are summarized in Table 1 at the last row in each panel. They are asymptotically unbiased, and comparable to the estimator of $\gamma$ under the fixed sample design (the last column and the last row of each panel).

When the alternative hypothesis is correctly specified in Table 2, $\delta = \theta$, the proposed tests achieve power similar to that from the fixed sample design, when there is no effect for the nuisance covariate ($\gamma = 0$). The power of the usual logrank test or Wald test based on the fixed sample designs can be much lower than the expected one, when the effect of the nuisance covariate increases, even when the fixed sample design relies on a correctly specified treatment effect, $\theta$ (Lagakos and Schoenfeld, 1984). In contrast, the proposed adaptive design incorporating the nuisance covariate can improve the power and maintain it at 90% level. In general, the bias-adjusted estimators have much smaller bias compared to the naive
estimators across all scenarios, and the asymptotic confidence interval at the final analysis maintains a coverage probability of \((1 - \alpha)\). As expected, it is important to adjust for the nuisance covariate, when the effect \(\gamma\) is large (relative to \(\theta\)). The covariate-adjusted estimators often have somewhat less bias than the estimators ignoring the nuisance covariate, and a larger block size also leads to a smaller bias. When the effect of the nuisance covariate is minor or moderate, the statistics based on the logrank and the Cox model perform similarly. We also calculate the naive estimator of \(\theta\) with or without adjusting for the covariate under the fixed sample design. The results are in accord with the early findings of Gail and others (1984) and Struthers and Kalbfleisch (1986) that the estimator of \(\theta\) can be biased with omitted covariates from the proportional hazards model under the fixed sample design. When ignoring the nuisance covariate that has a strong covariate effect (with \(\gamma = 1.2\)) in the adaptive design, the proposed estimator
Table 3. Empirical power and average number of events (ANE), bias-adjusted estimator and its 95% confidence interval for adaptive design with $B = 10$, and one-sided $\alpha = 0.025$, $1 - \beta = 0.9$, and alternative hypothesis $H_1: \delta = 0.693$ with a baseline hazard of 0.2.

<table>
<thead>
<tr>
<th></th>
<th>Logrank</th>
<th></th>
<th>Cox PH model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adaptive</td>
<td>OBF</td>
<td>Fixed sample</td>
<td>Adaptive</td>
</tr>
<tr>
<td>Data are generated from $\theta = 0.60$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma = 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical power</td>
<td>0.887</td>
<td>0.798</td>
<td>0.803</td>
<td>0.883</td>
</tr>
<tr>
<td>ANE</td>
<td>118.95</td>
<td>65.23</td>
<td>91.72</td>
<td>117.71</td>
</tr>
<tr>
<td>Adj. Est, $\theta$</td>
<td>0.619</td>
<td></td>
<td></td>
<td>0.622</td>
</tr>
<tr>
<td>Adj. 95% CP, $\theta$</td>
<td>0.931</td>
<td></td>
<td></td>
<td>0.931</td>
</tr>
<tr>
<td>Naive Est, $\theta$</td>
<td>0.679</td>
<td>0.676</td>
<td>0.610</td>
<td>0.681</td>
</tr>
<tr>
<td>Naive Est, $\gamma$</td>
<td></td>
<td>0.007</td>
<td>-0.006</td>
<td>0.008</td>
</tr>
<tr>
<td>$\gamma = 0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical power</td>
<td>0.864</td>
<td>0.690</td>
<td>0.733</td>
<td>0.866</td>
</tr>
<tr>
<td>ANE</td>
<td>108.53</td>
<td>76.07</td>
<td>76.13</td>
<td>108.44</td>
</tr>
<tr>
<td>Adj. Est, $\theta$</td>
<td>0.620</td>
<td></td>
<td></td>
<td>0.624</td>
</tr>
<tr>
<td>Adj. 95% CP, $\theta$</td>
<td>0.933</td>
<td></td>
<td></td>
<td>0.930</td>
</tr>
<tr>
<td>Naive Est, $\theta$</td>
<td>0.675</td>
<td>0.651</td>
<td>0.598</td>
<td>0.679</td>
</tr>
<tr>
<td>Naive Est, $\gamma$</td>
<td></td>
<td>0.501</td>
<td>0.510</td>
<td>0.506</td>
</tr>
<tr>
<td>$\gamma = 0.8$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical power</td>
<td>0.812</td>
<td>0.627</td>
<td>0.700</td>
<td>0.825</td>
</tr>
<tr>
<td>ANE</td>
<td>102.83</td>
<td>60.22</td>
<td>69.27</td>
<td>99.85</td>
</tr>
<tr>
<td>Adj. Est</td>
<td>0.612</td>
<td></td>
<td></td>
<td>0.622</td>
</tr>
<tr>
<td>Adj. 95% CP</td>
<td>0.905</td>
<td></td>
<td></td>
<td>0.912</td>
</tr>
<tr>
<td>Naive Est, $\theta$</td>
<td>0.673</td>
<td>0.629</td>
<td>0.594</td>
<td>0.687</td>
</tr>
<tr>
<td>Naive Est, $\gamma$</td>
<td></td>
<td>0.811</td>
<td>0.818</td>
<td>0.813</td>
</tr>
<tr>
<td>$\gamma = 1.2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical power</td>
<td>0.756</td>
<td>0.584</td>
<td>0.616</td>
<td>0.812</td>
</tr>
<tr>
<td>ANE</td>
<td>99.61</td>
<td>57.23</td>
<td>61.97</td>
<td>98.10</td>
</tr>
<tr>
<td>Adj. Est</td>
<td>0.580</td>
<td></td>
<td></td>
<td>0.609</td>
</tr>
<tr>
<td>Adj. 95% CP</td>
<td>0.910</td>
<td></td>
<td></td>
<td>0.927</td>
</tr>
<tr>
<td>Naive Est, $\theta$</td>
<td>0.636</td>
<td>0.621</td>
<td>0.579</td>
<td>0.664</td>
</tr>
<tr>
<td>Naive Est, $\gamma$</td>
<td></td>
<td>1.230</td>
<td>1.211</td>
<td>1.233</td>
</tr>
</tbody>
</table>

PH, proportional hazards; Adj. Est, adjusted estimator; CP, coverage probability.

of $\theta$ can be biased due to the misspecified model. The magnitude of the biases is comparable to that of the naive estimator of $\theta$ in the corresponding fixed sample design omitting the nuisance covariate.

As suggested by the editors, we included additional comparisons of power and estimation between the adaptive design and a classical group sequential design with fixed sample sizes. With the same type I and type II error rates as in the fixed sample designs, we planned a maximum number of five interim analyses, with roughly equal numbers of events observed between two consecutive analyses for the O’Brien & Fleming (OBF) design in Table 3. The results were summarized under the column “OBF” for this design. Table 3 shows the results when data are generated with a smaller log-hazard ratio than the assumed one ($\theta = 0.60$). It is clear that both logrank-based and Cox-model-based test statistics of the adaptive design have a substantial gain in power compared with the fixed sample tests with or without interim analyses. This is not surprising because the classical group sequential designs do not allow to modify the maximum
sample size. In addition, because of the interim efficacy tests and potential early termination of the trials, OBF designs have smaller average number of events (and power) compared to the fixed sample design without interim analyses. Similar to the naive estimator of $\theta$ under the adaptive design, the naive estimator of $\theta$ under OBF is also overestimated. For example, when $\gamma = 0.5$ and $\theta = 0.6$, the two naive estimators are 0.679 and 0.650 under Cox model for the adaptive and OBF designs. As expected, the naive estimators for $\gamma$ under both adaptive and OBF designs are unbiased when $Z$ and $W$ are independent.

When the effect of the nuisance covariate is large (relative to $\theta$), the Cox-model-based test statistic leads to an improved power compared to the test statistic, ignoring the nuisance covariate in both the adaptive and fixed sample designs. Moreover, the coverage probability of the proposed confidence interval at the final analysis is closer to the significance level with the nuisance covariate adjustment than without the adjustment. In general, the bias-adjusted estimators have smaller bias compared to the naive estimators across all scenarios. The naive estimator of $\theta$ under the fixed sample design in all tables is unbiased, and the coverage probabilities under the correctly specified models should be close to the nominal level, because there is no sequential sampling involved in the fixed sample design. However, the power of testing can be lower than expected when the log-hazard ratio is overestimated under the fixed sample design.

4. Example

To illustrate the method, we use data from a completed, multicenter randomized clinical trial comparing treatments for colon cancer. We show what would have happened in terms of inference and estimation, had the trial been adaptively designed. In the trial, patients with resected stages B and C colorectal carcinoma were randomized to receive either placebo, adjuvant levamisole (lev), or adjuvant levamisole plus fluorouracil (lev + 5-FU) for 1 year after surgery. One of the primary goals was to evaluate the efficacy of levamisole, either alone or in combination with 5-FU, to reduce recurrence of the disease. Detailed analyses of the trial are given by Moertel and others (1990) and Laurie and others (1989). For purposes of illustration, we restrict our attention to the comparison of efficacy between the placebo arm and the lev+5-FU arm. The variable of interest is time to recurrence of the disease or death, whichever occurs first. As was observed in the study, the presence of metastatic lymph nodes at the diagnosis is a determinant of disease recurrence and death (Moertel and others, 1990). Therefore, we incorporate the baseline lymph node information while monitoring the trial.

Among a total of 619 patients who entered the trial between the years of 1984 and 1987, we removed 12 patients from our analysis due to a lack of information about their lymph nodes. Accrual in the actual trial lasted from 1984 to 1987, with a median follow-up of 7.5 years. Using the usual Cox proportional hazards model in the fixed sample design with all observed data, the estimated log-hazard ratio (relative risk) is 0.49 (statistically significant with a one-sided $p$-value < 0.001) for disease-free survival between patients in the placebo arm and the lev+5-FU arm. The variable of interest is time to recurrence of the disease or death, whichever occurs first. As was observed in the study, the presence of metastatic lymph nodes at the diagnosis is a determinant of disease recurrence and death (Moertel and others, 1990). Therefore, we incorporate the baseline lymph node information while monitoring the trial.

Suppose that the investigators would like to detect a 50% decrease of the hazard rate in the treatment group, compared with the placebo group, in which patients have a baseline hazard rate of 0.23, assuming an exponential distribution. These assumptions are reasonably close to the protocol for a fixed sample design (Laurie and others, 1989) and the observed data. Given an accrual duration of 4 years and an additional 5 years of follow-up, a sample size of 619 would have about 80% power to detect a 50% decrease in the hazard rate of recurrence and/or death.

With a fixed accrual rate and additional follow-up, we assume that the accrual duration can be extended from 2 to 4 years, if necessary. We apply the proposed adaptive design procedure with the block size of 20 at each look (except $B_1 = 50$); the trial would be terminated with the null hypothesis being rejected after observing 90 events, i.e., upon completion of the third interim analysis. In this case, a total of 448 patients were entered to the study, and the total study duration was 5 years. Specifically, the trial would be stopped earlier under the adaptive design than with the fixed sample design and would claim the efficacy...
Adaptive design

of levamisole together with 5-FU, compared with the placebo. The adjusted log-relative risk is estimated to be 0.52 with 95% confidence interval of (0.06, 0.98). The inference based on the self-designing trial results in a test statistic of 2.2 with a \( p \)-value of 0.01. The naive estimate of the log-relative risk between the two treatment arms is 0.71 in the Cox proportional hazards model at the end of the third interim analysis, which is substantially overestimated compared to the one estimated at the end of the trial (0.49) under the fixed sample design.

5. DISCUSSION

We have presented a general inference and estimation theory for the self-designing trials with censored survival data under the commonly used semiparametric proportional hazards model. Shen and Cai (2003) generalized the self-designing trial from immediately observed outcomes to censored survival outcomes with staggered entry. However, the attention is still restricted to testing the equality of two survival distributions. A recent presentation by Yang and others (2005) also described a special case for testing, which is a generalization for the test statistic proposed by Shen and Cai (2003) to take nuisance covariates into consideration. The primary objective of this study is to develop a general estimation method after the inference in a self-designing trial with staggered entry survival data. In particular, at the end of a self-designing clinical trial, it is of great interest to estimate the relative hazard ratio between the two treatment arms and its corresponding confidence interval with or without nuisance covariates.

As demonstrated in the literature for other sequential design or in the fixed sample design settings (Lakagos and Schoenfeld, 1984; Tsiatis and others, 1985; Jennison and Turnbull, 1997), as well as in our study, the use of a test ignoring nuisance covariates often results in a loss of power when the nuisance covariate effects are strong, even though the fixed sample design is based on a correctly specified treatment effect. The use of covariate-adjusted tests at the interim analyses as well as at the final analysis of the adaptive design leads to a moderate gain in power and some improvement in efficiency for the estimates. The empirical results in our study show that the estimator of the treatment effect \( \theta > 0 \) is asymptotically biased toward zero if the nuisance covariate is omitted from the model in the adaptive design. This is analogous to the findings in Struthers and Kalbfleisch (1986) under the fixed sample design. While the effect of a nuisance covariate is mild or moderate compared to the primary treatment effect, the overall performance of the inference using the Cox model is similar to the inference using a logrank-type of test (omitting the nuisance covariates) under both adaptive design and fixed sample design.

The asymptotic properties for the sequential test statistics with the adjustment of nuisance covariates are mainly based on the work of Gu and Ying (1995). The weighted score process for testing the treatment effect while adjusting for other covariates in the Cox model has asymptotic normal distribution under the null hypothesis. Similarly, the distribution property for the proposed pivot has been derived by the corresponding variance–covariance matrix for all covariates. The basic results lead to the estimation of the treatment effect and its confidence interval at the end of an adaptive trial. The accuracy of the approximation under a variety of scenarios is fairly good.

It is often of interest to estimate the effect of nuisance covariates on the primary outcome at the end of an adaptive trial. Intuitively, the naïve estimators at the end of the trial for nuisance covariates should be unbiased when randomizing patients to different treatment arms is independent of the nuisance covariates. Because the interim decision rules which induce the bias into the estimation of the treatment effect at the end of the adaptive trial are devised to monitor the treatment effect only, rather than the nuisance covariates, the naïve estimate for the nuisance effect should be unbiased. The empirical results, as shown in the tables, confirm this intuition. On the other hand, the validity of the estimation procedure for \( \theta \) is still maintained when \( Z \) and \( W \) are not independent, since trial monitoring and subsequent decisions of trial conduct are not based on testing the nuisance covariates in the design. However, the naïve estimator of \( \gamma \) at the end of the study can be biased in this case.
ACKNOWLEDGMENTS

The authors thank the Associate Editor and a referee for helpful suggestions. This research was partially supported by National Cancer Institute Grant 2R01 CA079466. The research was done while the second author was visiting the Department of Biostatistics and Applied Mathematics, M. D. Anderson Cancer Center. Conflict of Interest: None declared.

APPENDIX A

A.1 Proof of $t_1, \ldots, t_m$ being stopping times

As suggested by one reviewer, the asymptotically independent increment structure of $U(t_j, \cdot)$ can be derived if $t_1, \ldots, t_j, \ldots, t_m$ are stopping times.

By definition (e.g. Fleming and Harrington, 1991, p 52), a stopping time $t_j$ for the filtration, $\{\mathcal{F}_t: t \geq 0\}$, is defined as a nonnegative random variable with the property that, for each $t \geq 0$, the event $\{t_j \leq t\}$ lies in $\mathcal{F}_t$. Suppose for the counting processes defined on Section 2.1, $\mathcal{F}_t = \sigma \{X_i(u), \Delta_i(u), Z_i, W_i, 0 \leq u \leq t, i = 1, \ldots\}$. The filtration $\mathcal{F}_{t_j}$ at time $t_j$ is the $\sigma$-field generated by the cumulated data up to the $j$th look, where $j = 1, 2, \ldots, m$. Intuitively, the time $t_j$ is a stopping time if at any fixed time $t$ one can observe whether or not $B_j$ events have already occurred, where $B_j$ is prefixed (Andersen and others, 1995, p 61). Thus, it suffices to show that $t_m$ is a stopping time, where $m$ is a random integer determined by the data observed up to the $(m-1)$th step.

Proving that $t_m$ is a stopping time is equivalent to showing that $\{t_m \leq t\} \in \mathcal{F}_t$ for any $t > 0$. By decomposing the event $\{t_m \leq t\}$ by $\{m = k\}$, we reexpress the event $\{t_m \leq t\}$ as

$$\{t_m \leq t\} = \bigcup_{k=1}^{\infty} (\{t_m \leq t\} \cap \{m = k\}) = \bigcup_{k=1}^{\infty} (\{t_k \leq t\} \cap (\{m = k\} \cap \{t_{k-1} \leq t\}))].$$

(A.1)

The last equality holds because $t_1 < t_2 < \cdots < t_{k-1} < t_k < \cdots$. Furthermore

$$\{t_k \leq t\} = \left\{ \text{Total number of events occurred in } [0, t] \geq \sum_{i=1}^{k} B_i \right\}$$

$$= \left\{ \sum_{i=1}^{n_k} N_i(u, t) \geq \sum_{i=1}^{k} B_i \right\} \in \mathcal{F}_t,$$

(A.2)

where $N_i(u, t) = I(X_i(u) \leq t, \Delta_i(u) = 1)$ and $u \leq t$. Recall that $\{m = k\} \in \mathcal{F}_{t_{k-1}}$ from Shen and Fisher (1999), which implies

$$\{m = k\} \cap \{t_{k-1} \leq t\} \in \mathcal{F}_t.$$  

(A.3)

Combining (A.1) through (A.3), $t_m$ is proved to be a stopping time. For any fixed $j$ ($<m$), it is just a special case of the above derivations. Therefore, $t_1, \ldots, t_j, \ldots, t_m$ are stopping times.

APPENDIX B

B.1 Derivation of the asymptotic distribution of $V_m(W)$

Following the same arguments as in Gu and Ying (1995), $n_j^{-1/2}U(t_j, \theta_0, \hat{\gamma}_j)$ defined in Section 2.3 can be decomposed as

$$n_j^{-1/2}U(t_j, \theta_0, \hat{\gamma}_j) = n_j^{-1/2}U(t_j, \theta_0, \gamma_0) + n_j^{-1/2}I_{zw}(t_j)I_{w_0w}(t_j)H(t_j, \theta_0, \gamma_0) + o_p(1),$$
where both $U(t_j, \theta_0, \gamma_0)$ and $H(t_j, \theta_0, \gamma_0)$ can be expressed as the sum of i.i.d. martingale integrals with independent increments. Thus, $n_j^{-1/2}U(t_j, \theta_0, \gamma_j)$ converges to a Gaussian process with independent increments and variance–covariance function $\sigma^2_z(t_j)$ under the null hypothesis. Under the alternatives, $n_j^{-1/2}U(t_j, \theta)$ behaves like a Gaussian process asymptotically with mean value $n_j\theta\sigma^2_z(t_j)$ and variance $\sigma^2_z(t_j)$. To have a more precise mathematical interpretation, $n_j^{-1/2}U(t_j, \theta)$ can be normalized under a sequence of local alternatives of $\theta_n = n^{-1/2}\theta_1$, and converges to $B(\sigma_z(t_j)) + \sigma_z(t_j)\gamma_1$, where $B$ is the standard Brownian motion process. By the independent increment property of the score process $\gamma(t)$, $U(t_j, \theta_0, \gamma_j)$ converges to an application of the theorem of Shen and Fisher (1999). Hence, the limit of $S(t_j, \gamma)$, which, by Slutsky’s theorem, can be written as

$$S(t_j, \gamma_j) = \frac{U(t_j, \theta_0, \gamma_j) - U(t_j-1, \theta_0, \gamma_j-1)}{n_j(\sigma^2_z(t_j) - \sigma^2_z(t_j-1))} + o_p(1),$$

where $o_p(1)$ is uniform in $t_j$. This, together with the independent increment property of $U(t_j, \theta_0, \gamma_j)$, leads to the asymptotic independence of $S(t_j, \gamma_j)$ and $S(t_k, \gamma_k)$ (for $j \neq k$). Under the contiguous alternatives, the limiting distribution of $S(t_j, \gamma_j)$ is normal with variance 1 and mean $\theta_1(\sigma^2_z(t_j) - \sigma^2_z(t_j-1))^{1/2}$. Hence, the limit of $S(t_j, \gamma_j)$ converges weakly to a zero-mean Gaussian process. Thus, asymptotic normality of $V_m(W)$ follows from an application of the theorem of Shen and Fisher (1999). Since the limiting distribution property of $V_m$ in Section 2.2 is a special case of $V_m(W)$ when $\gamma = 0$, we do not repeat that proof.

REFERENCES


[Received December 5, 2005; revised April 7, 2006; accepted for publication June 16, 2006]