A General Theory on Stochastic Curtailment for Censored Survival Data

D. Y. Lin, Q. Yao, and Zhiliang Ying

Stochastic curtailment is a valuable tool in monitoring long-term medical studies. Under this approach, one calculates the conditional power, which is the probability of rejecting the null hypothesis at the scheduled end of the study given the existing data at the interim analysis, along with certain speculation about the future data. The conditional power may be used to aid the decision to terminate a study prematurely or to extend a study beyond its originally planned duration. This article provides a formal and systematic investigation into the use of stochastic curtailment in the context of censored survival data. To enhance generality, we introduce a broad class of statistics that includes two-sample weighted log-rank statistics, as well as the partial likelihood score statistic for testing treatment difference with covariate adjustment under the proportional hazards model. We establish the weak convergence under both the null hypothesis and contiguous alternatives for this class of statistics when calculated repeatedly over the calendar time (i.e., time of interim analysis). Further, we derive the conditional distributions of these statistics calculated at the end of the study given all the data collected up to the interim look or given the statistics calculated at the interim look, and provide analytic expressions for the corresponding conditional powers. These results enable us to address several subtle issues involved in the definition and implementation of conditional power for censored survival data, especially when there is staggered patient entry with a potential time trend in the survival distribution, when the Gehan-type weight function is used, or when treatment is not independent of covariates. For randomized clinical trials, we show that very simple formulas can be used to calculate the conditional powers of the unweighted log-rank test (with or without covariate adjustment) under both the null and alternative hypotheses. Simulation studies demonstrate that the conditional powers for survival studies can be accurately evaluated through the proposed formulas even when the sample size is small. An illustration with data taken from a colon cancer study is provided.

KEY WORDS: Brownian motion; Clinical trial; Conditional power; Proportional hazards; Sequential analysis; Weighted log-rank test.

1. INTRODUCTION

For ethical, scientific, and economic reasons, it is customary to periodically examine the accumulating data in a long-term medical study so that the study may be terminated when there is sufficient evidence either for or against the null hypothesis. The main statistical tools to aid this decision-making process are group sequential tests (Lan and DeMets 1983; O’Brien and Fleming 1979; Pocock 1977) and stochastically curtailed tests (Lan, Simon, and Halperin 1982). Under the group sequential approach, one performs a significance test at each interim look using a fraction of the overall type I error and terminates the study whenever the null hypothesis is rejected. Under the stochastic curtailment approach, one calculates the so-called conditional power—the conditional probability of rejecting the null hypothesis at the scheduled end of the study given the current data, along with some speculation about the future data—and may terminate the study when the conditional power is very high or very low. The conditional power may also aid the decision to extend a study beyond its originally planned duration of accrual/follow-up.

Lan et al. (1982) and Lan and Wittes (1988, sec. 2) developed a simple and elegant theory on stochastic curtailment for Brownian motions. Suppose that \( B(t) \) (\( 0 \leq t \leq 1 \)) is a Brownian motion with linear drift \( \theta \). Suppose also that the one-sided fixed-sample–size test \( B(1) \) for testing \( H_0: \theta = 0 \) has type I error \( \alpha \) and power \( 1 - \alpha' \) at \( H_1: \theta = \theta_1 \) (\( \theta_1 > 0 \)).

It follows from well-known properties of Brownian motion that the conditional distribution of \( B(1) \) given \( B(t) \) is normal with mean \( B(t) + \theta(1 - t) \) and variance \( 1 - t \). Thus the conditional probability that \( B(1) > z_\alpha \) given \( B(t) \) is

\[
\rho(t; \theta) = 1 - \Phi \left( \frac{z_\alpha - B(t) - \theta(1 - t)}{(1 - t)^{1/2}} \right),
\]

where \( \Phi \) is the standard normal cumulative distribution function and \( \Phi(z_\alpha) = 1 - \alpha \). Furthermore, if one stops the study to reject \( H_0 \) when \( \rho(t; 0) > \rho_0 \) and stops to reject \( H_1 \) when \( \rho(t; \theta_1) < 1 - \rho_1 \), then

\[
\begin{align*}
&\text{type I error probability } \leq \alpha / \rho_0, \\
&\text{type II error probability } \leq \alpha' / \rho_1.
\end{align*}
\]

Because of the central limit theorem, formulas (1) and (2) are applicable to many one-sample and two-sample statistics with immediate responses (Lan et al. 1982; Lan and Wittes 1988).

The most common response variable in long-term medical studies is the (possibly censored) survival time. To test the equality of two survival distributions with censored data, one commonly uses weighted log-rank statistics (Andersen, Borgan, Gill, and Keiding 1993, chap. V; Fleming and Harrington 1991, chap. 7). Sequential tests based on these statistics have been studied extensively (e.g., Gu and Lai 1991; Jones and Whitehead 1979; Slud and Wei 1982; Tsai 1982). In addition, Tsai, Rosner, and Trichler (1985) and Gu and Ying (1995) have developed sequential tests based on the partial likelihood score statistic for testing treatment difference with covariate adjustment under the proportional hazards model. When the allocation of

© 1999 American Statistical Association
Journal of the American Statistical Association
June 1999, Vol. 94, No. 446, Theory and Methods
treatment is not made at random, it is essential to adjust for baseline prognostic factors in testing treatment difference. For randomized studies, the covariate adjustment may be used to improve efficiency, to account for stratified designs, to correct for chance imbalance in prognostic variables, and to adjust for the covariates related to the entry time and/or censoring time.

Some special cases of stochastically curtailed tests with censored survival data have been studied in the statistical literature and utilized in important clinical trials. Lan and Wittes (1988, sec. 4) suggested that the conditional power of the log-rank test under the null hypothesis of no treatment difference could be evaluated by formula (1) and illustrated this calculation with the Beta-Blocker Heart Attack Trial (BHAT). In the interim analysis conducted 9 months before the scheduled closing date of BHAT, the normalized log-rank statistic was 2.82, which exceeded the critical value of 2.23 specified by the O’Brien–Fleming sequential method. Furthermore, the conditional powers under the null trend were found to range from .80 for 120 additional deaths to .94 for 60 additional deaths, with a power of .89 for the best estimate of 80 additional deaths. Thus not only were the existing results statistically significant, but also it was considered unlikely that the results of the remaining 9 months would reverse any conclusion about the observed treatment effect. These statistical considerations, along with some medical considerations, led to the termination of BHAT 9 months ahead of schedule (DeMets, Hardy, Friedman, and Lan 1984). Lan and Wittes (1988, sec. 4) also mentioned that the conditional power of the log-rank test under an alternative hypothesis was difficult to calculate but might be evaluated by simulation “if one is willing to postulate a proportional hazards model or otherwise parameterize the survival curves.”

Pawitan and Hallstrom (1990) studied the conditional power for the log-rank test under the proportional hazards model by assuming that the death rate is low and censoring is independent of treatment, and provided an application to the Cardiac Arrhythmia Suppression Trial (CAST). At the second interim analysis of CAST, the normalized log-rank statistic was −3.22, which crossed the lower advisory boundary for adverse effect. The conditional powers evaluated at the prespecified alternative hypothesis of hazard ratio of .7 were .23 for the realistic projection of 300 total deaths and .58 even with the utilization of 425 deaths. The study was terminated prematurely, due to the existing evidence of adverse effect and the low conditional power of establishing beneficial effect. The early termination of CAST has had a tremendous negative impact on the clinical investigations of anti-arrhythmia drugs.

Andersen (1987) considered conditional power calculations for the (unweighted) log-rank test in a model with constant hazard function and censoring, and described how such calculations were used to determine whether or not to extend an alcoholic cirrhosis clinical trial beyond its predesigned end of the trial. The study was terminated as scheduled with a negative conclusion, because the conditional power for detecting a clinically relevant benefit would have been very low even if the study had been extended for 2 years.

Henderson, Fisher, Weber, Hammermeister, and Sethi (1991) described a simulation method to evaluate the conditional powers of the log-rank and Gehan–Wilcoxon tests for a Veterans Administration study. The study was extended for 5 years to provide sufficient power to detect treatment benefit, although the existing evidence was not strong enough.

There does not exist any analytic result for a weighted log-rank test with a nonconstant weight function or for the partial likelihood score test with covariate adjustment. Even for the two-sample (unweighted) log-rank test, it is currently unknown whether or not the analytic results of Andersen (1987), Lan and Wittes (1988), and Pawitan and Hallstrom (1990) hold under broader conditions. Therefore, it is highly desirable and useful to conduct a formal and systematic investigation into the use of stochastic curtailment with censored survival data.

In this article we develop a general theory on stochastic curtailment for censored survival data, providing analytic expressions to evaluate the conditional powers under both the null and alternative hypotheses for all the aforementioned tests (i.e., weighted log-rank tests and covariate-adjusted partial likelihood score test). We discuss a number of subtle issues pertaining to the use of stochastic curtailment in the context of censored survival data and ascertain the conditions under which formulas (1) and (2) are valid. In the special case of the two-sample (unweighted) log-rank test considered by the previous authors, we show that it is possible to calculate the conditional powers analytically under both the null and alternative hypotheses without imposing the stringent assumptions of Andersen (1987) or Pawitan and Hallstrom (1990).

To present our results for weighted log-rank statistics and covariate-adjusted partial likelihood score statistic in a compact form, we introduce a class of weighted log-rank–type statistics for testing no treatment effect with possible covariate adjustment under the proportional hazards model. This class of statistics not only includes the aforementioned two types of statistics as special cases, but also provides additional flexibilities for detecting nonproportional hazards treatment effect with covariate adjustment.

2. THEORY AND METHODS

2.1 A General Class of Statistics

Let T be the survival time, Z be the treatment indicator, and W be a p × 1 vector of covariates. We allow possible dependence between Z and W. Suppose that the hazard function of T conditional on Z and W is

$$
\lambda(x|Z, W) = \lambda_0(x)e^{\beta_0 q_0(x)Z + \gamma_0 w},
$$

(3)

where $\lambda_0(x)$ is an arbitrary baseline hazard function, $q_0(x)$ is a deterministic function, and $\beta_0$ and $\gamma_0$ are unknown regression parameters. Model (3) implies that for any w,
\[
\log \frac{\lambda(x | Z = 1, W = w)}{\lambda(x | Z = 0, W = w)} \propto q_0(x).
\]

Of course, model (3) reduces to the standard proportional hazards model if \(q_0 = 1\).

Let \(R\) and \(C\) be the entry time and censoring time; that is, the times when the subject enters and withdraws from the study. Both \(T\) and \(C\) are measured from \(R\), whereas \(R\) is measured in calendar time. We assume that \(W\) is bounded and \(T\) is independent of \(R\) and \(C\) conditional on \(Z\) and \(W\). Suppose that \(n\) independent subjects will be enrolled into the study eventually. When the data are reviewed at calendar time \(t\), the observation time and censoring indicator for the \(i\)th subject are \(X_i(t) = T_i \wedge C_i \wedge (T - R_i)^+\) and \(\Delta_i(t) = I\{T_i \leq C_i \wedge (T - R_i)\}\). Here and in the sequel, \(a \wedge b = \min(a, b), a \vee b = \max(a, b), a^+ = a \vee 0, \text{ and } I\{\cdot\}\) is the indicator function. For any fixed \(t\), the available data \(\{X_i(t), \Delta_i(t), Z_i, W_i\} (i = 1, \ldots, n)\) are no different from the usual survival data. It is necessary to index the data by \(t\), because the observation time and censoring indicator evolve over calendar time \(t\).

Let \(N_i(t, x) = \Delta_i(t)I\{X_i(t) \leq x\}\) and \(Y_i(t, x) = I\{X_i(t) \geq x\}\). Note that \(N_i(t, \cdot)\) and \(Y_i(t, \cdot)\) are the familiar counting process and at-risk process for the \(i\)th subject (Fleming and Harrington 1991, chap. 0) based on the data available at calendar time \(t\). Under the null hypothesis \(H_0: \beta_0 = 0\), the partial likelihood score functions for \(\beta_0\) and \(\gamma_0\) at calendar time \(t\) are

\[
\dot{l}_\beta(t; \gamma) = \sum_{i=1}^n \int_0^t q_0(x) \{Z_i - E_Z(t, x, \gamma)\} N_i(t, dx)
\]

and

\[
\dot{l}_\gamma(t; \gamma) = \sum_{i=1}^n \int_0^t \{W_i - E_W(t, x, \gamma)\} N_i(t, dx),
\]

where

\[
E_Z(t, x, \gamma) = \frac{\sum_{j=1}^n Y_j(t, x)e^{\gamma W_j}Z_j}{nS^{(0)}(t, x, \gamma)},
\]

\[
E_W(t, x, \gamma) = \frac{\sum_{j=1}^n Y_j(t, x)e^{\gamma W_j}W_j}{nS^{(0)}(t, x, \gamma)},
\]

and \(S^{(0)}(t, x, \gamma) = n^{-1}\sum_{j=1}^n Y_j(t, x)e^{\gamma W_j}\). Here and in the sequel,

\[
\int_0^t H(t, x)N_i(t, dx) = \Delta_i(t)I\{X_i(t) \leq s\}H(t, X_i(t)).
\]

The restricted maximum partial likelihood estimator (MLE) of \(\gamma_0\), denoted by \(\hat{\gamma}_t\), is the solution to \(\{l_\gamma(t; \hat{\gamma}_t) = 0\}\). The corresponding score statistic for testing \(H_0\) is \(l_\beta(t; \hat{\gamma}_t)\).

In practice, the function \(q_0(x)\) is unknown, and thus \(l_\beta(t; \hat{\gamma}_t)\) may not be realized. Instead, we propose the following class of statistics:

\[
U(t; \hat{\gamma}_t) = \sum_{i=1}^n \int_0^t Q(t, x)\{Z_i - E_Z(t, x, \hat{\gamma}_t)\} N_i(t, dx),
\]

where the random weight function \(Q(t, x)\) is constructed from the data available at time \(t\). It is natural to let \(Q\) mimic the weight functions of two-sample weighted log-rank statistics. Assume that \(Q(t, x)\) converges in probability to \(q(t, x)\) uniformly in \(t\) and \(x\). If \(q\) is free of \(t\), then we abbreviate \(q(t, x)\) as \(q(x)\). Of course, \(U\) evolves over \(t\). The argument \(\hat{\gamma}_t\) in \(U\) will often be suppressed.

The class of statistics given in (4) encompasses the existing censored data statistics mentioned in Section 1. If \(Q = 1\), then \(U\) becomes the partial likelihood score statistic for testing \(H_0: \beta_0 = 0\) under the proportional hazards model \(\lambda(x | Z, W) = \lambda_0(x)e^{\beta_0 Z + \gamma_0 W}\). When no covariates are involved (i.e., \(p = 0\)), \(U\) reduces to two-sample weighted log-rank statistics. Thus \(U\) may be called weighted log-rank statistics with covariate adjustment. We refer to \(U\) as the (unweighted) log-rank statistic if \(Q = 1\), whether or not covariates are involved. The most popular nonconstant weight functions are those of Peto–Prentice–Wilcoxon and Gehan–Wilcoxon. For the Gehan–Wilcoxon–type statistic, \(Q(t, x)\) may be either \(n^{-1}\sum_{i=1}^n Y_i(t, x)\) or \(S^{(0)}(t, x, \hat{\gamma}_t)\), and \(q(t, x)\) depends on both \(t\) and \(x\). There are two possible generalizations of the Peto–Prentice–Wilcoxon weight function to accommodate covariates: one is the Kaplan–Meier estimator for the marginal survival function of \(T\) based on \(\{X_i(t), \Delta_i(t)\} (i = 1, \ldots, n)\), and the other is the baseline survival function estimator \(e^{-\lambda_0(t, x)}\), where

\[
\hat{\lambda}_0(t, x) = \sum_{i=1}^n \int_0^x \frac{N_i(t, du)}{nS^{(0)}(t, u, \hat{\gamma}_t)},
\]

which is the Breslow estimator of the cumulative baseline hazard function \(\lambda_0(x) = \int_0^x \lambda_0(u) du\) calculated at time \(t\).

The limit of the marginal survival function estimator will be free of \(t\) if \(R\) is independent of \(W\), whereas that of the baseline survival function estimator is always free of \(t\).

### 2.2 Asymptotic Properties of \(U(t)\)

The null distribution of \(U(t)\) was mentioned by Biliama, Gu, and Ying (1997, cor. 5.1), though no detail was provided on the motivation or utilization of \(U(t)\) with a nonconstant weight function. Our Theorem 1 given here extends the result of Biliama et al. (1997) by showing the weak convergence of \(U(t)\) under the sequence of contiguous alternatives:

\[
H_1^{(n)}: \lambda^{(n)}(x | Z, W) = \lambda_0(x)e^{\gamma^{(n)}(x)Z + \gamma_0 W},
\]

where \(\gamma^{(n)} = n^{-1/2p} b\) and \(b\) is a fixed real number. We introduce some notation before stating the theorem. Let

\[
E_{ZW}(t, x, \gamma) = \frac{\sum_{j=1}^n Y_j(t, x)e^{\gamma W_j}Z_j W_j}{nS^{(0)}(t, x, \gamma)},
\]

and

\[
E_{WW}(t, x, \gamma) = \frac{\sum_{j=1}^n Y_j(t, x)e^{\gamma W_j}W_j^{\otimes 2}}{nS^{(0)}(t, x, \gamma)},
\]

where \(a^{\otimes 2} = aa\). Also, let \(s^{(0)}(t, x), e_Z(t, x), e_W(t, x), e_{ZW}(t, x), \text{ and } e_{WW}(t, x)\) be the limits of \(S^{(0)}(t, x, \gamma_0)\),
Finally, let
\[
G(t, s) = \int_0^{t \wedge s} q(t, x)q(s, x)\{e_Z(t \wedge s, x) - e_Z(t \wedge s, x)^2\} \times s(0)(t \wedge s, x) d\Lambda_0(x),
\]
\[
D(t, s) = \int_0^{t \wedge s} q(t \wedge s, x)\{e_Z(t \wedge s, x)
- e_Z(t \wedge s, x)e_{W}(t \wedge s, x)\} s(0)(t \wedge s, x) d\Lambda_0(x),
\]
\[
\Omega(t) = \int_0^t \{e_{WW}(t, x) - e_{WW}(t, x)^2\} s(0)(t, x) d\Lambda_0(x),
\]
\[
G(t) = G(t, t), \text{ and } D(t) = D(t, t). \ \tilde{G}(t) \text{ is obtained from } G(t) \text{ by replacing } q(t, x)^2 \text{ with } q(t, x)q_0(x), \text{ and } D_0(t) \text{ is } D(t) \text{ evaluated at } q(t, x) = q_0(x).
\]

**Theorem 1.** Under \( H_1^{(n)} \), the process \( n^{-1/2}U(t) \) converges weakly to a Gaussian process with mean \( b \tilde{C}(t) \) and with covariance function \( \xi(t, s) \) at \( (t, s) \), where \( \xi(t, s) = \tilde{G}(t) - D(t) \Omega^{-1}(t) D_0(t) \) and \( \Omega(t, s) = G(t, s) - D(t \wedge s) \Omega^{-1}(t \wedge s) D(t, s) \).

Theorem 1, proved in the Appendix, is essential to the power and efficiency considerations. This theorem contains the null distribution of \( U \) as a special case: under \( H_0, n^{-1/2}U(t) \) converges weakly to a zero-mean Gaussian process with covariance function \( \xi(t, s) \). If \( q(t, x) \) is independent of \( t \), then \( G(t, s) = G(t \wedge s) \) and \( D(t, s) = D(t \wedge s) \), implying that \( \xi(t, s) = \xi(t \wedge s), \text{ where } \xi(t, t) = \xi(t, t). \) In this case, \( n^{-1/2}U(t) \) is asymptotically a Brownian motion in the time scale of \( \xi(t) \) under \( H_0 \). Furthermore, if \( q(t, x) = q_0(x), \text{ then } \xi(t, s) = \xi(t, t) \), which means that \( n^{-1/2}U(t) \) is asymptotically a Brownian motion with linear drift \( b \) in the time scale of \( \xi(t) \) under \( H_1^{(n)} \). On the other hand, if \( q(t, x) = q_0(x) \neq q_0(x) \), then the asymptotic distribution of \( n^{-1/2}U(t) \) has a nonlinear drift for nonzero \( b \). If covariates are not involved or are independent of treatment, then \( \xi(t) = \tilde{G}(t) \) and \( \xi(t, s) = G(t, s) \).

We can estimate \( \xi(t, s) \) or \( \xi(t) \) consistently by replacing the unknown parameters in \( G, D, \) and \( \Omega \) with their respective sample estimators. Denote the resulting estimators of \( \xi(t, s) \) and \( \xi(t) \) by \( \hat{\xi}(t, s) \) and \( \hat{\xi}(t) \). Also, denote \( n\hat{\xi}(t) \) by \( V(t) \). Thus to test \( H_0: \beta_0 = 0 \) at calendar time \( t \), we refer \( S(t) = U(t)/V(t) \) to the standard normal distribution.

The test statistic \( S(t) \) is always valid (at least asymptotically) in terms of preserving the type I error probability regardless of what the weight function \( Q(t, x) \) is, as long as the set of covariates \( W \) satisfies the proportional hazards model (if covariates are involved at all). The power, however, depends critically on the choice of \( Q \), as evident from Theorem 1. It is widely recognized that the optimal limiting weight function for the conventional two-sample weighted log-rank statistic is proportional to the true log hazard ratio of the two treatments (Andersen et al. 1993, VIII.2.3; Fleming and Harrington 1991, sec. 7.4). The following theorem, proved in the Appendix, shows that this result continues to hold even with covariate adjustment.

**Theorem 2.** If \( Q(t, x) \) is so specified that \( q(t, x) = q_0(x) \), then \( S(t) \) is most powerful not only within the proposed class of statistics but also among all possible semi-parametric statistics under the sequence of contiguous alternatives \( H_1^{(n)} \).

This theorem pertains to the optimality of \( S(t) \) for each fixed calendar time \( t \). It remains an open question as to what an optimal sequence of weight functions should be when \( S(t) \) is calculated repeatedly over the course of a study.

### 2.3 Stochastic Curtailment

The key to the implementation of stochastic curtailment lies in the evaluation of conditional power. In theory, one may define the conditional power as the probability of rejecting \( H_0 \) at the scheduled end of the study (i.e., calendar time \( \tau \)) given any aspect of the data collected up to calendar time \( t (t < \tau) \) along with certain speculation about the future data. If the sequential statistic \( B(t) \) (asymptotically) a Brownian motion (with a linear drift) in \( t \) or in some transformation of \( t \), then the conditional power for \( B(\tau) \) given \( B(t) \) can be evaluated by (1). In our setting, the asymptotic distribution of \( n^{-1/2}U(t) \) will not be a Brownian motion if \( q(t, x) \) depends on \( t \). In their seminal work, Lan et al. (1982) defined the conditional power as being conditional on all of the data collected up to \( t \). This is also how conditional power is commonly interpreted by practitioners. Under this definition, it may be a nontrivial task to calculate the conditional power even if \( B(t) \) is a Brownian motion. The reason is that the conditional distribution of \( B(1) \) or \( B(\tau) \) given all the data available at time \( t \) is generally not the same as that given \( B(t) \). In our setting, it turns out to be important to distinguish between these two types of conditional distributions/powers. On the other hand, the inequalities in (2) will hold as long as the conditional power is defined on an increasing sequence of \( \sigma \) fields (Lan et al. 1982, thms. 1 and 2).

We first consider the conditional power of \( S(\tau) \) given \( S(t) \). The following result follows immediately from Theorem 1.

**Corollary 1.** For \( \beta_0 = O(n^{-1/2}) \), the conditional distribution of \( S(\tau) \) given \( S(t) \) or \( U(t) \) is asymptotically normal with mean
\[
\mu(t; \beta_0) = S(t) \frac{\xi(t, \tau)}{\xi^{1/2}(t)\xi^{1/2}(\tau)} + n^{1/2}\beta_0\left(\xi(t, \tau) - \xi(t, t)\xi^{-1}(t)\xi(t)\right) \xi^{1/2}(\tau)
\]
and variance \( \sigma^2(t) \equiv 1 - \xi^2(t, \tau)/\xi(t, \tau)\xi(t)\).

In view of Corollary 1, the conditional power of \( S(\tau) \) given \( S(t) \) is asymptotically
\[
\rho(t; \beta_0) = 1 - \Phi \left\{ c - \frac{\mu(t; \beta_0)}{\sigma(t)} \right\}, \tag{6}
\]
assuming that \( H_0: \beta_0 = 0 \) is to be rejected at \( \tau \) iff \( S(\tau) > c \). If \( q(t, x) = q(x) \) for all \( t \), then \( \xi(t, \tau) = \xi(t) \), which implies
that
\[ \mu(t; \beta_0) = S(t)^{1/2} + \frac{n^{1/2} \beta_0 \{ \xi(\tau) - \bar{\xi}(t) \}}{\xi^{1/2}(\tau)} \] (7)
and
\[ \sigma^2(t) = 1 - \tilde{t}, \] (8)
where \( \tilde{t} = \xi(t)/\xi(\tau) \). Furthermore, if \( q(x) = q_0(x) \), then \( \xi(t) = \xi(t) \) for all \( t \) and, consequently,
\[ \mu(t; \beta_0) = S(t)^{1/2} + \beta_0 \{ n \xi(\tau) \}^{1/2} (1 - \tilde{t}). \] (9)

The case of \( q(x) = q_0(x) \) fits into the Brownian motion framework described in Section 1, with \( \tilde{t}, S(t)^{1/2} \), and \( \beta_0 \{ n \xi(\tau) \}^{1/2} \) playing the roles of \( t, B(t), \) and \( \theta \); thus (1) is applicable. Formula (1) can also be used to calculate the conditional power under \( H_0 \) if \( q(t, x) = q(x) \neq q_0(x) \). In all other cases, (6), together with appropriate expressions for \( \mu \) and \( \sigma \), must be used.

If \( U(t) \) has independent increments (asymptotically)—that is, \( q(t, x) = q(x) \) for all \( t \)—then the conditional distribution of \( S(\tau) \) given \( U(t) \) is equivalent to that given \( \{ U(s); s \leq t \} \). Under this condition, the analysis time \( t \) may depend on the past history of \( U(t) \) and its variation process, and the upper bounds of type I and II error rates given in (2) apply to \( \rho(t; \beta_0) \) of (6). However, \( U(t) \) may have independent increments. An important example is the Gehan–Wilcoxon–type statistic. Without the independent increment structure, Corollary 1 itself holds only when \( t \) is completely independent of the data, and inequalities of (2) may not hold.

To circumvent the potential unpleasantness of dependent increments, we propose a slightly different conditioning. Define
\[ U_\tau(t) = \sum_{i=1}^{n} \int_{0}^{t} q(\tau, x) \{ Z_i - E_Z(t, \tau, \gamma_i) \} N_i(t, dx), \]
which is obtained from (4) by replacing \( Q(t, x) \) with \( q(\tau, x) \). Because of the convergence of \( Q \) to \( q, n^{-1/2} U_\tau(\tau) \) is asymptotically the same as \( n^{-1/2} U_\tau \), although the asymptotic equivalence does not hold for \( t < \tau \) unless \( q(t, x) = q(x) \) (for all \( t \)). It then follows from Theorem 1 that, under \( H_1^{(n)}, n^{-1/2} U_\tau(t) \) converges weakly to a Gaussian martingale (i.e., Gaussian process with independent increments) with mean function \( b_\tau(t) \equiv b \{ G_\tau(t) - D_\tau(t) \} \Omega^{-1}(t) D_0(t) \} \) and variance function \( \xi_\tau(t) \equiv G_\tau(t) - D_\tau(t) \} \Omega^{-1}(t) D_0(t) \} \), where \( G_\tau(t), G_\tau(t), \) and \( D_\tau(t) \) are obtained from \( G(t), G(t), \) and \( D(t) \) by replacing \( q(t, x) \) with \( q(\tau, x) \). Thus we have the following result.

Corollary 2. For \( \beta_0 = O(n^{-1/2}) \), the conditional distribution of \( S(\tau) \) given \( \{ U_\tau(s); s \leq t \} \) is asymptotically normal with mean
\[ \mu_\tau(t; \beta_0) = n^{-1/2} U_\tau(t) + n^{1/2} \beta_0 \{ \xi(\tau) - \bar{\xi}(t) \} \] \[ \xi^{1/2}(\tau) \]
and variance \( \sigma^2_\tau(t) \equiv 1 - \xi_\tau(t)/\xi(\tau) \).

To calculate the conditional power of \( S(\tau) \) given \( \{ U_\tau(s); s \leq t \} \), we replace \( \mu(t; \beta_0) \) and \( \sigma(t) \) in (6) by \( \mu_\tau(t; \beta_0) \) and \( \sigma_\tau(t) \). Denote the resulting power function by \( p_\tau(t; \beta_0) \). Under \( H_0; \beta_0 = 0, \rho_\tau(t; \beta_0) \) is equivalent to \( p(t; 0) \) of (1), with \( n^{-1/2} U_\tau(t)/\xi^{1/2}(\tau) \) and \( \xi_\tau(t)/\xi(\tau) \) playing the roles of \( B(t) \) and \( t \). Formula (1) cannot be used for nonzero \( \beta_0 \) unless \( q(t, x) = q_0(x) \) for all \( t \). Because \( U_\tau(t) \) has independent increments, it is valid to calculate the conditional power at any calendar time \( t \), which may depend on \( \{ U_\tau(s); s \leq t \} \) and its variation process. Furthermore, the inequalities in (2) hold. Of course, \( p(t; \beta_0) = p(t; 0) \) if \( q(t, x) = q(x) \) for all \( t \). It is important to note that without the condition of \( q(t, x) = q(x), U_\tau(t) (t < \tau) \) is not a statistic in the conventional sense, because \( q(\tau, x) \) involved in \( U_\tau(t) \) cannot be estimated from the data available at time \( t \).

We now turn to the more delicate problem of calculating the conditional power of \( S(\tau) \) given all of the data collected up to \( t \), denoted by \( F_t \). The conditional distribution of \( S(\tau) \) given \( F_t \) cannot be derived from Theorem 1. The following theorem is proved in the Appendix.

Theorem 3. For \( \beta_0 = O(n^{-1/2}) \), the conditional distribution of \( S(\tau) \) given \( F_t \) is asymptotically normal with mean
\[ \mu_F(t; \beta_0) = n^{-1/2} U_\tau(t) + n^{1/2} \beta_0 \{ \xi(\tau) - \bar{\xi}(t) \} \] \[ \xi^{1/2}(\tau) \]
and variance \( \sigma^2_F(t) \equiv 1 - \xi_F(t)/\xi(\tau) \), where
\[ m(t) = \sum_{i=1}^{n} \int_{0}^{t} \left[ q(\tau, x) \{ Z_i - E_Z(t, \tau, \gamma_i) \} \right] \]
\[ - D'(\tau) \Omega^{-1}(\tau) [ W_i - E_W(\tau, x) ] \]
\[ \times \{ N_i(t, dx) - Y_i(t, x) e^{\beta_0 q_0(x)Z_i + \xi_F W_i} d\Lambda_0(x) \}, \]
\[ \xi_F(t) = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} \mathcal{E} \left( \int_{0}^{t} \left[ q(\tau, x) \{ Z_i - E_Z(t, \tau, \gamma_i) \} \right] \right. 
\[ - D'(\tau) \Omega^{-1}(\tau) [ W_i - E_W(\tau, x) ] \} \]
\[ \times \left. Y_i(t, x) e^{\beta_0 q_0(x)Z_i + \xi_F W_i} d\Lambda_0(x) \right), \]
and \( \mathcal{E} \) denotes the expectation. Furthermore, if entry time is independent of all other random variables, then
\[ \mu_F(t; \beta_0) = n^{-1/2} U_\tau(t; \gamma_0) + n^{1/2} \beta_0 \{ \xi(\tau) - \bar{\xi}(t) \} \] \[ \xi^{1/2}(\tau) \]
and
\[ \xi_F(t) = G_\tau(t) + D'(\tau) \Omega^{-1}(\tau) [ \Omega(\tau) \Omega^{-1}(\tau) D(\tau) \] \[ - 2D'(\tau) \Omega^{-1}(\tau) D_\tau(t), \] (10)
where
\[ U_F(t; \gamma) = \sum_{i=1}^{n} \int_{0}^{t} \left[ q(\tau, x) \{ Z_i - E_Z(t, \tau, \gamma_i) \} \right] \]
\[ - D'(\tau) \Omega^{-1}(\tau) [ W_i - E_W(t, x, \gamma) ] N_i(t, dx), \]
and
\[ \tilde{\xi}_\tau(t) = \tilde{G}_\tau(t) - D'(\tau) \Omega^{-1}(\tau) D_0(t). \]

If there is a time trend in patients’ characteristics or in the standards of medical care, then entry time will not be independent of all other random variables. Except for the second part of Theorem 3, our theory does not require the independence of entry time to all other random variables, although this independence is required by virtually all of the existing methods. In general, formula (1) cannot be used to calculate the conditional power of \( S(\tau) \) given \( \mathcal{F}_t \). Instead, one should use (6) by replacing \( \mu \) and \( \sigma \) with \( \mu_\tau \) and \( \sigma_\tau \). Denote the resulting power function by \( \rho_\tau \). The analysis time \( t \) may depend on any past history of the data, and the inequalities of (2) hold. If covariates are absent or independent of treatment, then \( D(\tau) = 0 \), which greatly simplifies the forms of \( m, \xi_\tau, \mu_\tau \), and \( \xi_\tau \).

Comparisons of the conditional means and variances among Corollaries 1 and 2 and Theorem 3 reveal that in general the conditional distribution of \( S(\tau) \) given \( S(t) \) or \( U(t) \) is not the same as that given \( \mathcal{F}_t \), even when \( n^{-1/2} U(t) \) is a Brownian motion. The reason for this discrepancy is that a Brownian motion or Gaussian martingale is automatically Markovian with respect to the filtration (i.e., history) generated by the process itself, but not necessarily with respect to the filtration generated by all of the data up to time \( t \); the latter generally being larger than the former. In our setting, the data filtration contains the information about \( \beta_0 \) as well as the information about nuisance functions (i.e., \( \gamma_0, \Lambda_0 \), and entry time/censoring distributions), whereas the self-generated filtration of \( U(t) \) or \( U_r(t) \) comprises the part of the data used to make inference about \( \beta_0 \).

Let us consider randomized clinical trials. For such studies, treatment is independent of covariates, which implies that \( D(t) = D_r(t) = 0 \) (for all \( t \)). Suppose that entry time is independent of all other random variables. It then follows from Corollary 2 and the second part of Theorem 3 that \( \xi_\tau(t) = \xi_\tau(t) = G_r(t), \xi_\tau(t) = \xi_\tau(t) = G_r(t), \) and \( \mu_\tau(t; \beta_0) \) is asymptotically equivalent to \( \mu_r(t; \beta_0) \). Suppose now that entry time is not independent of all other random variables, but that the treatment allocation is made completely at random, so that entry time is independent of treatment. Then it is easy to see that \( e_\tau(t, x) \) will be free of treatment if censoring is independent of treatment or if censoring is independent of anything else but treatment (i.e., \( R, T \) and \( W \)). By Corollary 2 and the first part of Theorem 3, if \( e_\tau(t, x) \) is free of \( t \) and \( D(t) = D_r(t) = 0 \) (for all \( t \)), then \( \xi_\tau(t) = \xi_\tau(t) = G_r(t) \) and \( \mu_\tau(t; \beta_0) \) is asymptotically equivalent to \( \mu_r(t; \beta_0) \). In short, \( \rho_\tau \) and \( \rho_r \) are equivalent for randomized clinical trials under mild conditions on entry time/censoring.

Which type of conditional power should be used? The undesirable features of \( \rho(t; \beta_0) \) when \( q(t, x) \neq q(x) \) suggest the use of \( \rho_r(t; \beta_0) \) rather than \( \rho(t; \beta_0) \) or the selection of \( Q(t, x) \) such as \( q(t, x) = q(x) \) (for all \( t \)). In the latter case, \( \rho_r \) and \( \rho \) are equivalent. Thus the real choice is between \( \rho_r \) and \( \rho_\tau \). The purpose of using the conditional power is to predict the event \( \{ S(\tau) > c \} \). This prediction can be made more accurate by conditioning on the entire \( \sigma \) filtration \( \mathcal{F}_t \) rather than on the smaller filtration \( \{ U_r(s); s \leq t \} \). In fact, it is not difficult to show that
\[
E[I\{S(\tau) > c\} - \rho_\tau(t; \beta_0)]^2 
\leq E[I\{S(\tau) > c\} - \rho_r(t; \beta_0)]^2. \tag{11}
\]

As discussed in the preceding paragraph, \( \rho_r \) and \( \rho_\tau \) are equivalent in randomized studies under mild conditions on entry time/censoring. This equivalence further supports the use of \( \rho_r \) rather than \( \rho \) when \( q(t, x) \) depends on \( t \). When \( \rho_r \) and \( \rho_\tau \) are equivalent, one may use the simpler expression \( \rho_r \), and the analysis time \( t \) may depend on \( \mathcal{F}_t \). When they are not equivalent, \( \rho_\tau \) is more accurate in that inequality (11) is strict, although it is still reasonable to use \( \rho_r \) because \( U_r(t) \) contains all of the information in the current data that can be used to make inference about \( \beta_0 \).

How does one actually evaluate the conditional power? If \( q(t, x) = q(x) \) (for all \( t \)), then the conditional mean and variance for \( \rho \) or equivalently \( \rho_r \) are given in (7) and (8). In these two expressions, \( \xi(\tau) \) and \( \tilde{\xi}(\tau) \) are determined in the design stage, whereas \( \xi(t) \) and \( \tilde{\xi}(t) \) can be consistently estimated from the data at time \( t \); therefore, it is easy to evaluate the conditional power. If \( q(t, x) \neq q(x) \), then one should use \( \rho_r \) rather than \( \rho \), as suggested previously. A slight complication in using \( \rho_r \) when \( q(t, x) \neq q(x) \) is that \( U_r, \tilde{\xi}(t) \), and \( \xi(t) \) in the conditional mean and variance expressions involve \( \tau \). Unlike \( q(t, x) \), \( q(\tau, x) \) is determined by the distributions of \( T, C \), and \( R \) specified in the design stage rather than being estimated from the data available at time \( t \). This complication provides an additional reason for not using a weight function whose limit depends on \( t \). When entry time is independent of all other random variables, the main difference between \( \rho_r \) and \( \rho_\tau \) is that the latter explicitly involves \( \gamma_0, D(\tau), \) and \( \Omega(\tau) \), which are determined by the design parameters; therefore, it is slightly more complicated to evaluate \( \rho_\tau \) than \( \rho_r \). Without the independence, an additional complication in evaluating \( \rho_r \) is that \( m(t) \) and \( \xi(t) \) involve \( e \xi(\tau, x) \), \( e \Omega(\tau, x) \), and \( \Lambda_0(t) \), which are again determined by the distributions of \( Z, W, T, C \), and \( R \) specified in the design stage. We expect \( \rho_\tau \) to be more sensitive to the specification of individual design parameters than \( \rho_r \).

As mentioned previously, for randomized trials in which entry time is independent of all other random variables or treatment is independent of entry time and censoring time, \( \rho_\tau \) is equivalent to \( \rho_r \), which can be easily implemented, especially under \( q(t, x) = q(x) \) (for all \( t \)).

In the foregoing discussion, we assumed that the quantities evaluated at \( \tau \), such as \( \xi(\tau), q(\tau, x) \), and \( e \xi(\tau, t) \), are determined in the design stage. These quantities are calculated by numerical integration or simulation once the distributions of \( W, T, R, T \), and \( C \) are specified. If the actual accrual, failure, and censoring patterns differ substantially from those anticipated in the design stage, then it may be necessary to reevaluate these quantities (again by numerical integration or simulation). In fact, one important use of stochastic curtailment is to aid the decision to extend a study beyond its originally planned duration when the patient accrual is lower than expected.
Let us take a closer look at the special case of the unweighted log-rank statistic (with or without covariate adjustment) for randomized clinical trials. For simplicity, assume that $Z$ is independent of $C$ and $R$. Then $\xi(t) \approx n^{-1} d(t) \sigma_Z^2$, where $d(t)$ is the total number of deaths observed up to time $t$ and $\sigma_Z^2$ is the variance of $Z$, which is $\frac{1}{2}$ if an equal number of patients is allocated to each of the two treatment groups (Tsai at. al. 1985). In addition, $\xi(t) \approx n^{-1} \sum_{i=1}^{n} \Delta_i (t) q_0 \{X_i(t) \sigma_Z^2\}$, which of course equals $\xi(t)$ if $q_0 = 1$. Given these two simple expressions, it is straightforward to calculate the conditional power via (6)–(8). If $q_0(x) = 1$, then one may also use (1) by setting $B(t) = S(t) \{d(t)/d(\tau)\}^{1/2}, \theta = \beta_0 \{d(t) \sigma_Z^2\}^{1/2}$, and $t = d(t)/d(\tau)$. This evaluation is extremely easy, because $S(t)$ can be obtained from any standard software, $d(t)$ and $\sigma_Z^2$ involve simple enumerations of the available data, and $d(\tau)$ is just the projected number of deaths observed by the scheduled closing date. Note that the conditions imposed here imply that $\rho_x = \rho_\tau = \rho$. These results not only justify the proposal of Lan and Wittes (1988) to calculate the conditional power under $H_0$ using (1), but also show that the conditional powers under proportional and nonproportional hazards alternatives can be calculated analytically without imposing overly stringent assumptions.

The simulation approach has some obvious drawbacks. First, calculating the conditional power by simulation is cumbersome and time-consuming, and requires special programming effort. Second, there is randomness in the simulated conditional power, and the result is not reproducible by others. In contrast, the analytic approach is convenient and provides a unique and reproducible answer. The simulation approach has one potential advantage over the analytic approach in that it does not rely on the asymptotic approximations and thus may be more reliable in small samples. But our extensive simulation studies revealed that the asymptotic approximations are very accurate even in fairly small samples.

Table 1 displays the results of our simulation studies for the important special case of the two-sample unweighted log-rank statistic. We considered total sample sizes of 100, 200, and 500, with half of the patients randomly assigned to each treatment. In each trial, a one-sided test with .05 significance level was scheduled to perform when a given number of deaths $d(\tau)$ was observed, and the conditional power was calculated at the interim analysis time $t$ at which there was $d(t)$ number of observed deaths. The entry times were generated from the uniform(0, 1) distribution; the survival times, from the exponential distribution with hazard rate $2 e^{0.2}$, where $Z$ was the treatment indicator. The survival data at the interim look $t$ were generated with either $\beta = 0$ or $\beta = 3$. For each interim dataset, the conditional power was evaluated through (1) with $z_0 = 1.645$, $B(t) = S(t) \{d(t)/d(\tau)\}^{1/2}, \theta = \beta_0 \{d(\tau)/4\}^{1/2}$, and $t = d(t)/d(\tau)$, where $S(t)$ is the normalized log-rank statistic calculated at time $t$, and $\beta_0 = 0$ under $H_0$ and $\beta_0 = 3$ under $H_1$. The conditional power was also evaluated by simulation, which generated the residual survival times for those who were still alive at $t$ from the exponential distribution with hazard rate $2 e^{0.2}$. The analytic and simulation-based conditional powers are shown by $\hat{\rho}$ and $\hat{\rho}$ in Table 1. As evident from the table, there are remarkably good agreements between the analytic and simulated conditional powers. The discrepancies are so small that they are more likely to have been caused by the randomness of simulation than by the inaccuracy of the asymptotic approximations. The simulation-based conditional powers shown in Table 1 are based on 10,000 replications. The estimates were much more unstable with 1,000 replications. In view of these results and the aforementioned drawbacks of the simulation approach, we

<table>
<thead>
<tr>
<th>$n$</th>
<th>$d(t)$</th>
<th>$d(\tau)$</th>
<th>$S(t)$</th>
<th>$\rho(t; \theta) = 0$</th>
<th>$\rho(t; \theta) = 3$</th>
<th>$\rho(t; \theta) = 0$</th>
<th>$\rho(t; \theta) = 3$</th>
<th>$\hat{\rho}(t; \theta) = 0$</th>
<th>$\hat{\rho}(t; \theta) = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>25</td>
<td>50</td>
<td>1.819</td>
<td>.036</td>
<td>.024</td>
<td>.596</td>
<td>.600</td>
<td>2.353</td>
<td>.510</td>
</tr>
<tr>
<td>200</td>
<td>50</td>
<td>100</td>
<td>1.817</td>
<td>.121</td>
<td>.124</td>
<td>.337</td>
<td>.340</td>
<td>2.815</td>
<td>.871</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>200</td>
<td>1.469</td>
<td>.196</td>
<td>.195</td>
<td>.581</td>
<td>.591</td>
<td>9.686</td>
<td>.360</td>
</tr>
<tr>
<td>100</td>
<td>150</td>
<td>.529</td>
<td>.246</td>
<td>.238</td>
<td>.646</td>
<td>.639</td>
<td>3.135</td>
<td>.943</td>
<td>.948</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>.641</td>
<td>.046</td>
<td>.046</td>
<td>.426</td>
<td>.433</td>
<td>2.414</td>
<td>.535</td>
<td>.525</td>
</tr>
<tr>
<td>500</td>
<td>300</td>
<td>.668</td>
<td>.026</td>
<td>.029</td>
<td>.343</td>
<td>.344</td>
<td>2.525</td>
<td>.778</td>
<td>.776</td>
</tr>
</tbody>
</table>

**Table 1. Formula-Based Conditional Powers $\rho$ and Simulation-Based Conditional Powers $\hat{\rho}$ for the Two-Sample Unweighted Log-Rank Tests**

$\mu(t; \theta) = 0$ and $\mu(t; \theta) = 3$ are the conditional powers. $\beta_0 = 0$ and $\beta_0 = 3$ based on (1), and $\hat{\rho}(t; \theta)$ and $\hat{\rho}(t; \theta) = 3$ are the corresponding estimates based on simulation of 10,000 replications.
recommend that the conditional power be calculated analytically rather than by simulation.

4. A REAL EXAMPLE

In a national intergroup trial to investigate the effectiveness of adjuvant therapy for resected colon carcinoma, 315, 310, and 304 patients with stage C disease were randomly assigned to observation, levamisole (Lev) alone, and levamisole combined with 5-FU (Lev + 5-FU) (Moertel et al., 1990). Enrollment of patients began in March 1984 and ended in October 1987. Survival was the primary endpoint. The investigators anticipated a total of 500 deaths in the three groups, which ensured that each pairwise comparison by a one-sided log-rank test with .05 significance level would have a power of about .863 to detect a hazard ratio of 1.35; that is, \( \beta_0 = .3 \).

A four-stage group sequential boundary of O'Brien and Fleming (1979) with an overall type I error rate of .05 was utilized. Analyses were planned to occur after approximately 125, 250, 375, and 500 deaths were observed. At the first analysis conducted in December 1987, there was little evidence of any treatment effect. At the time of the second analysis in September 1989, the value of the log-rank test for comparing Lev + 5-FU and observation crossed the stopping boundary, and the trial was terminated. There were 125 and 301 observed deaths at the two interim looks.

We now illustrate how stochastic curtailment might have been used to monitor this trial. Because prior to randomization the patients were stratified according to three important prognostic variables (i.e., the invasion of the primary lesion, the interval since surgery, and the number of lymph nodes involved), it is reasonable to adjust for these three variables in treatment comparisons. Table 2 shows the observed values of the log-rank tests for comparing Lev alone versus observation and Lev + 5-FU versus observation at the two interim looks. The conditional powers were evaluated according to (1) by setting \( z_\alpha = 1.645, B(t) = 5(t) \{(d(t)/d(\tau))^{1/2}, \theta = \beta_0 \{(2/3)d(\tau)/4\}^{1/2}, \) and \( t = d(t)/d(\tau) \), where \( \beta_0 = 0 \) under \( H_0 \) and \( \beta_0 = .3 \) under \( H_3 \). \( d(\tau) \) is the corresponding number at the scheduled closing date. We considered \( d(\tau) = 450, 500, \) and 550. The number of deaths observed by the scheduled closing date \( \tau \) depends on the actual accrual, survival, and censoring patterns, which may differ appreciably from what were anticipated in the design stage. The number of observed deaths for each treatment comparison is roughly two-thirds of the total number in the three groups. Although not shown in Table 2, \( (2/3)d(t)/4 \) agrees well with the actual variance of \( U(t) \) for each pairwise comparison. According to the simulation studies of Section 3, which were designed to mimic this study, the analytic conditional powers given in Table 2 are accurate.

As discussed in Section 1, it is possible to set up a formal stopping rule based on the conditional power instead of the group sequential test. If one stopped the trial to reject \( H_0: \beta_0 = 0 \) when \( \rho(t; 0) > .8 \) and stopped to “accept” \( H_0 \) when \( 1 - \rho(t; .3) > .8 \), the actual type I error rate would be no more than .0625, and the actual power would be at least .83 at \( \beta_0 = .3 \). At the first look, \( \rho(t; 0) \) and \( 1 - \rho(t; .3) \) are far less than .8, providing little evidence either for or against \( H_0 \).

The results are more interesting at the second look. Let us first compare Lev + 5-FU and observation. The conditional powers are quite high even under \( H_0 \). Without the covariate adjustment, \( \rho(t; 0) \) is less than the threshold of .8 if \( d(\tau) = 500 \) or 550 and greater than .8 if \( d(\tau) = 450 \); with the covariate adjustment, \( \rho(t; 0) \) is less than .8 if \( d(\tau) = 550 \) and greater than .8 if \( d(\tau) = 500 \) or 450. Thus whether or not the study should be terminated on the basis of stochastic curtailment depends on the number of deaths anticipated at the scheduled closing date and to a lesser extent on whether or not covariate adjustment is made. In practice, stochastic curtailment is not the only statistical consideration. As mentioned in Section 1, the decisions to terminate B'HAT and CAST took into account both the results of group sequential testing and stochastic curtailment. In this case the result of group sequential testing was not as overwhelming as in B'HAT or CAST, although the O'Brien–Fleming boundary (i.e., 2.582) was crossed. Thus the decision on whether or not to terminate the trial would depend on the anticipated number of deaths at the scheduled closing date and of course on medical considerations.

<table>
<thead>
<tr>
<th>( t )</th>
<th>Comparison</th>
<th>( d(\tau) = 450 )</th>
<th>( d(\tau) = 500 )</th>
<th>( d(\tau) = 550 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \rho(t; 0) )</td>
<td>( \rho(t; .3) )</td>
<td>( \rho(t; 0) )</td>
<td>( \rho(t; .3) )</td>
</tr>
<tr>
<td></td>
<td>Unadjusted</td>
<td>.710</td>
<td>.681</td>
<td>.069</td>
</tr>
<tr>
<td></td>
<td>Adjusted</td>
<td>.560</td>
<td>.506</td>
<td>.058</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.762</td>
<td>.811</td>
<td>.851</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.732</td>
<td>.787</td>
<td>.831</td>
</tr>
<tr>
<td>1.63</td>
<td>Unadjusted</td>
<td>1.12</td>
<td>.840</td>
<td>.107</td>
</tr>
<tr>
<td></td>
<td>Adjusted</td>
<td>1.23</td>
<td>.852</td>
<td>.900</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.110</td>
<td>.874</td>
<td>.908</td>
</tr>
<tr>
<td>2.726</td>
<td>Unadjusted</td>
<td>.004</td>
<td>.087</td>
<td>.007</td>
</tr>
<tr>
<td></td>
<td>Adjusted</td>
<td>.278</td>
<td>.166</td>
<td>.306</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.084</td>
<td>.191</td>
<td>.418</td>
</tr>
<tr>
<td>2.842</td>
<td>Unadjusted</td>
<td>2.726</td>
<td>.994</td>
<td>.710</td>
</tr>
<tr>
<td></td>
<td>Adjusted</td>
<td>2.842</td>
<td>.996</td>
<td>.994</td>
</tr>
</tbody>
</table>

As discussed in Section 1, it is possible to set up a formal stopping rule based on the conditional power instead of the group sequential test. If one stopped the trial to reject \( H_0: \beta_0 = 0 \) when \( \rho(t; 0) > .8 \) and stopped to “accept” \( H_0 \) when \( 1 - \rho(t; .3) > .8 \), the actual type I error rate would be no more than .0625, and the actual power would be at least .83 at \( \beta_0 = .3 \). At the first look, \( \rho(t; 0) \) and \( 1 - \rho(t; .3) \) are far less than .8, providing little evidence either for or against \( H_0 \).

The results are more interesting at the second look. Let us first compare Lev + 5-FU and observation. The conditional powers are quite high even under \( H_0 \). Without the covariate adjustment, \( \rho(t; 0) \) is less than the threshold of .8 if \( d(\tau) = 500 \) or 550 and greater than .8 if \( d(\tau) = 450 \); with the covariate adjustment, \( \rho(t; 0) \) is less than .8 if \( d(\tau) = 550 \) and greater than .8 if \( d(\tau) = 500 \) or 450. Thus whether or not the study should be terminated on the basis of stochastic curtailment depends on the number of deaths anticipated at the scheduled closing date and to a lesser extent on whether or not covariate adjustment is made. In practice, stochastic curtailment is not the only statistical consideration. As mentioned in Section 1, the decisions to terminate B'HAT and CAST took into account both the results of group sequential testing and stochastic curtailment. In this case the result of group sequential testing was not as overwhelming as in B'HAT or CAST, although the O'Brien–Fleming boundary (i.e., 2.582) was crossed. Thus the decision on whether or not to terminate the trial would depend on the anticipated number of deaths at the scheduled closing date and of course on medical considerations.

<table>
<thead>
<tr>
<th>( t )</th>
<th>Comparison</th>
<th>( d(\tau) = 450 )</th>
<th>( d(\tau) = 500 )</th>
<th>( d(\tau) = 550 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \rho(t; 0) )</td>
<td>( \rho(t; .3) )</td>
<td>( \rho(t; 0) )</td>
<td>( \rho(t; .3) )</td>
</tr>
<tr>
<td></td>
<td>Unadjusted</td>
<td>.710</td>
<td>.681</td>
<td>.069</td>
</tr>
<tr>
<td></td>
<td>Adjusted</td>
<td>.560</td>
<td>.506</td>
<td>.058</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.762</td>
<td>.811</td>
<td>.851</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.732</td>
<td>.787</td>
<td>.831</td>
</tr>
<tr>
<td>1.63</td>
<td>Unadjusted</td>
<td>1.12</td>
<td>.840</td>
<td>.107</td>
</tr>
<tr>
<td></td>
<td>Adjusted</td>
<td>1.23</td>
<td>.852</td>
<td>.900</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.110</td>
<td>.874</td>
<td>.908</td>
</tr>
<tr>
<td>2.726</td>
<td>Unadjusted</td>
<td>.004</td>
<td>.087</td>
<td>.007</td>
</tr>
<tr>
<td></td>
<td>Adjusted</td>
<td>.278</td>
<td>.166</td>
<td>.306</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.084</td>
<td>.191</td>
<td>.418</td>
</tr>
<tr>
<td>2.842</td>
<td>Unadjusted</td>
<td>2.726</td>
<td>.994</td>
<td>.710</td>
</tr>
<tr>
<td></td>
<td>Adjusted</td>
<td>2.842</td>
<td>.996</td>
<td>.994</td>
</tr>
</tbody>
</table>
For the comparison between Lev alone and observation, the conditional powers are very low even under $H_1: \beta_0 = .3$. Without the covariate adjustment, $\rho(t_2; 3)$ is less than the threshold of .2 if $d(\tau) = 450$ or 500 and greater than .2 if $d(\tau) = 550$; with the covariate adjustment, $\rho(t_2; 3)$ is less than .2 if $d(\tau) = 450$ and greater than .2 if $d(\tau)$ is 500 or 550. Thus it would not be worthwhile to continue the study if one anticipated about 450 observed deaths at the scheduled closing date, whereas it might be worthwhile to continue if one anticipated more than 500 deaths. By contrast, the group sequential testing suggested that the study be continued regardless of the anticipated number of deaths at the scheduled closing date, because the sequential boundary was not crossed. This demonstrates that stochastic curtailment is much more useful than group sequential testing in deciding whether or not to abandon a study due to the lack of power. Of course, it would be unethical to prolong the trial to assess the effectiveness of Lev alone if the existing evidence was judged to be strong enough to claim a beneficial effect of Lev + 5-FU.

5. DISCUSSION

During the monitoring process of accumulating data, a question naturally arises as to whether the trend in the current data is so impressive, either for or against the null hypothesis, that the rejection or acceptance of the null hypothesis by the final test is likely to have been determined. Stochastic curtailment provides a formal answer to this question and thus is quite appealing to medical investigators. This device may be used by itself or in conjunction with group sequential testing. As evident from Section 4, the stopping boundaries based on conditional power tend to be wider than those of group sequential tests especially when the number of interim looks is small, but this conservatism buys us the flexibility to perform frequent looks at unplanned or data-dependent times. It is particularly convenient to use stochastic curtailment to aid the decision to abandon a study when the conditional power appears poor. Stochastic curtailment may also aid the decision to extend a study beyond its originally planned duration.

The conditional power given the current data and the conditional power given the current statistic seem to have been used interchangeably in the literature. In fact, the former is the usual definition and interpretation, whereas the latter is what most of the existing theory and methods, including (1), have been developed for. Although immaterial for many one-sample and two-sample statistics with immediate responses, the distinction is important in certain situations, particularly when the sequential statistic does not have independent increments or when the score statistic for the primary parameter is not orthogonal to that of the nuisance parameter(s). Another subtlety is that (2) applies to the conditional power given the current data and the conditional power given the current value of any process that has independent increments, but not to the conditional power given the current value of a sequential statistic with dependent increments.

The developments in this article have filled several critical gaps of knowledge about stochastic curtailment for censored survival data. First, we formally justified the proposal of Lan and Wittes (1988) to use (1) to calculate the conditional power of the (unweighted) log-rank test under the null hypothesis of no treatment difference, and also showed that the analytic evaluation for this test is straightforward under both the proportional and nonproportional hazards alternatives. Second, we provided analytic expressions to calculate the conditional powers of any weighted log-rank statistic under both the null and alternative hypotheses, and determined the conditions under which (1) is applicable as well as the conditions under which the conditional power given the current value of the statistic is equal to that given the current data. Third, we demonstrated that it is possible, although trickier, to calculate analytically the conditional power for a sequential statistic that does not have Brownian motion properties, such as the Gehan–Wilcoxon statistic. Simulation methods were used to calculate the conditional powers of Gehan–Wilcoxon-type statistics for the Veterans Administration study described by Henderson et al. (1991) and for the recently terminated Antiarrhythmics Versus Implantable Defibrillators trial (AVI Investigators 1997); the analytic formulas presented in Section 2 could have been used instead. Fourth, we dealt with the important and subtle issue of covariate adjustment especially when covariates are not independent of treatment. Although covariate adjustment is not essential for randomized clinical trials, Meier (1983, p. 180) pointed out that “a major exclusion is the common situation in which allocation to treatment is not made at random, so that statistical comparability of treatment groups cannot be assumed. In such studies, extensive adjustment for baseline variables is not simply an option with merits and limitations to be weighed, but an absolute necessity.” The proposed theory enables one to use stochastic curtailment in nonrandomized studies. For nonrandomized studies, early termination is often driven more by economic and scientific reasons than by ethical considerations.

We derived the conditional power under contiguous alternatives. This was for theoretical convenience and should not be regarded as a limitation of the proposed methods. Whenever one relies on asymptotic arguments to study the power of a test statistic, it is only meaningful to consider contiguous alternatives. In applications, one can choose any value of $\beta_0$ that is not excessively large. Simulation results of Section 3 demonstrated that the asymptotic approximations derived under contiguous alternatives yield accurate evaluation of the conditional power under fixed alternatives.

We have focused on testing the null hypothesis of no treatment difference; that is, $H_0: \beta_0 = 0$. It is straightforward to modify the formulas of Section 2 to test the null hypothesis of a non-zero $\beta_0$. There are some key differences though. First, $q_0(x)$ must now be assumed to be known even under $H_0$. Second, it is essential to adjust for all the covariates which affect entry time and/or censoring time even in
randomized studies. Third, $\rho_F$ and $\rho_r$ are equivalent only for randomized clinical trials in which entry time is independent of everything else.

The class of statistics studied in this article generalizes the conventional two-sample weighted log-rank statistics by allowing the adjustment of covariates. It also extends the familiar partial likelihood score statistic for testing no treatment difference in the presence of covariates by incorporating a weight function that can be chosen to yield asymptotically efficient tests against alternatives of various forms. This class of statistics can be used in both fixed-sample-size and sequential studies. We have not discussed repeated significance testing for the proposed class of statistics, because many of its special members have been extensively studied (Gu and Ying 1995; Slud and Wei 1982; Tsiatis 1982; Tsiatis et al. 1985), and the same principles apply to the remaining cases. As discussed in Section 2, if the limiting weight function $q(t, x)$ does not depend on $t$, then, under $H_0$, the process $n^{-1/2}U(t)$ is asymptotically a Brownian motion on the time scale of $\xi(t)$, in which case the error-spending function approach of Lan and DeMets (1983) can be used. Without the Brownian motion structure, one must resort to the approach of Slud and Wei (1982).

The validity of $S(t)$ depends on correct modeling of the effects of $W$ on $T$. For randomized studies, it is possible to modify the variance estimator of $U(t)$ so that $S(t)$ will remain valid even when model (3) is incorrect. Specifically, by combining the arguments of Lin and Wei (1989) with those given in the proof of Theorem 1, we can show that $U(t)$ is centered at 0 under the null hypothesis of no treatment difference regardless of the true model for $W$. In addition, a robust variance estimator for $U(t)$ is

$$\hat{\Omega}(t; \gamma) = n^{-1} \sum_{i=1}^{n} \int_{0}^{t} \{E_{\gamma}W(t, x, \gamma) - E_{\gamma}W(t, x, \gamma)^{\otimes 2}\} N_i(t, dx),$$

and $\gamma^*$ is on the line segment between $\gamma_1$ and $\gamma_0$. Define

$$M_i^{(n)}(t, x) = N_i(t, x) - \int_{0}^{x} \{e^{\gamma(t, u)} - 1\} \gamma_0 e^{\gamma_i(t, u)} d\Lambda_0(u). \quad (A.2)$$

For each $t$, $M_i^{(n)}(t, x) (i = 1, \ldots, n)$ are martingales with respect to the $\sigma$-filtration $\mathcal{F}_x \equiv \sigma\{N_i(t, u), Y_i(t, u), \mathcal{W}; u \leq x, i = 1, \ldots, n\}$, which consists of the data observed by calendar time $t$ and failure time $x$. We may express (A.1) as

$$U_Z(t; \gamma_1) = n^{-1/2}U_Z(t) - \hat{D}^*(t; \gamma^*)\hat{\Omega}^{-1}(t; \gamma^*)n^{-1/2}U_{\gamma}(t),$$

$$+ n^{-1/2} \sum_{i=1}^{n} \int_{0}^{t} Q(t, x) \{Z_i - E_Z(t, x, \gamma_0)\} \times \{e^{\gamma(t, u)} - 1\} \gamma_0^{\gamma(i(t, u))} d\Lambda_0(x)$$

$$- \hat{D}^*(t; \gamma^*)\hat{\Omega}^{-1}(t; \gamma^*) \times n^{-1/2} \sum_{i=1}^{n} \int_{0}^{t} \{W_i - E_{\gamma}W(t, x, \gamma_0)\} \{e^{\gamma(t, u)} - 1\} \gamma_0^{\gamma(i(t, u))} d\Lambda_0(x), \quad (A.3)$$

where

$$U_Z(t) = \sum_{i=1}^{n} \int_{0}^{t} Q(t, x) \{Z_i - E_Z(t, x, \gamma_0)\} M_i^{(n)}(t, dx)$$

and

$$U_{\gamma}(t) = \sum_{i=1}^{n} \int_{0}^{t} \{W_i - E_{\gamma}W(t, x, \gamma_0)\} M_i^{(n)}(t, dx).$$

By the counting-process martingale arguments (Andersen et al. 1993, chaps. II and VII), $\hat{D}(t; \gamma^*)$ and $\hat{\Omega}(t; \gamma^*)$ converge in probability to $D(t)$ and $\Omega(t)$, respectively. Therefore,

$$U_Z(t) = \sum_{i=1}^{n} \int_{0}^{t} q(t, x) \{Z_i - e_Z(t, x)\} M_i^{(n)}(t, dx) + o_p(n^{1/2})$$

and

$$U_{\gamma}(t) = \sum_{i=1}^{n} \int_{0}^{t} \{W_i - e_{\gamma}W(t, x)\} M_i^{(n)}(t, dx) + o_p(n^{1/2}).$$

By the Taylor series expansion, the last two terms on the right
side of (A.3) become
\[ n^{-1/2} \beta^{(n)} \left[ \int_0^t Q(t, x) q_0(x) \right. \\
\times \left\{ E_Z(t, x, \gamma_0) - E_Z(t, x, \gamma_0)^2 \right\} S^{(0)}(t, x, \gamma_0) d\Lambda_0(x) \\
- \hat{D}'(t; \gamma^*) \Omega^{-1}(t; \gamma^*) \int_0^t q_0(x) \\
\times \left\{ E_Z W(t, x, \gamma_0) - E_Z(t, x, \gamma_0) E W(t, x, \gamma_0) \right\} \\
\times S^{(0)}(t, x, \gamma_0) d\Lambda_0(x) \left. \right] + o_p(1), \]
which converges in probability to \( b \xi(t) \). Thus (A.3) can be written as
\[ n^{-1/2} U(t; \hat{\gamma}_i) \]
\[ = n^{-1/2} \sum_{i=1}^n \int_0^t [q(t, x) \{ Z_i - e_Z(t, x) \} \\
- D'(t) \Omega^{-1}(t) \{ W_i - e W(t, x) \} ] M_i^{(n)}(t, dx) \\
+ b \xi(t) + o_p(1). \] (A.4)
It then follows from the classical multivariate central limit theorem, together with straightforward (although somewhat tedious) covariance calculations, that \( n^{-1/2} U(t; \hat{\gamma}_i) \) converges in finite-dimensional distributions to a Gaussian process with mean \( b \xi(t) \) and covariance function \( \xi(t, s) \). The weak convergence then follows from the tightness of \( U_Z(t) \) and \( U_W(t) \), which can be verified via the techniques of Billingsley (1977).

**Proof of Theorem 2**

Note that
\[ \xi(t) = \lim_{n \to \infty} n^{-1} \sum_{i=1}^n E \left( \int_0^t [q(t, x) \{ Z_i - e_Z(t, x) \} \\
- D'(t) \Omega^{-1}(t) \{ W_i - e W(t, x) \} ] M_i^{(n)}(t, dx) \\
\times \int_0^t [q_0(x) \{ Z_i - e_Z(t, x) \} \\
- D'_0(t) \Omega^{-1}(t) \{ W_i - e W(t, x) \} ] M_i^{(n)}(t, dx) \right), \]
and
\[ \xi(t) = \lim_{n \to \infty} n^{-1} \sum_{i=1}^n E \left( \int_0^t [q(t, x) \{ Z_i - e_Z(t, x) \} \\
- D'(t) \Omega^{-1}(t) \{ W_i - e W(t, x) \} ] M_i^{(n)}(t, dx) \right)^2, \]
where \( E \) denotes the expectation. It then becomes apparent by the Cauchy–Schwarz inequality that \( \xi(t) \leq \{ \xi(t) \xi_0(t) \}^{1/2} \), where \( \xi_0(t) = \xi(t) \) evaluated at \( q(t, x) = q_0(x) \). Thus the noncentrality parameter \( b \xi(t) / \xi^{1/2}(t) \) or the power of \( S(t) \) is maximized at \( q(t, x) = q_0(x) \), in which case \( \xi(t) = \xi(t) = \xi(t) \).

To prove that \( S(t) \) with \( q(t, x) = q_0(x) \) is the most powerful test among all possible semiparametric tests under \( H_i^{(n)} \), it suffices to show that \( S(t) \) is asymptotically efficient (in terms of the Pitman efficiency) under a parametric submodel of (3). Let us consider the one-dimensional parametric submodel
\[ \lambda(x|Z, W; t) = \lambda(x) \exp(\beta q_0(x) \{ Z - e_Z(t, x) \}) \\
+ \gamma_0 W - \beta D_0(t) \Omega^{-1}(t) \{ W - e_W(t, x) \}, \]
where only \( \beta \) is regarded as an unknown parameter. For such a parametric model, the likelihood score test is known to be asymptotically efficient. The score statistic for testing \( H_0: \beta = 0 \) can be written as
\[ U_0(t) = \sum_{i=1}^n \int_0^t [q_0(x) \{ Z_i - e_Z(t, x) \} \\
- D'_0(t) \Omega^{-1}(t) \{ W_i - e_W(t, x) \} ] M_i(t, dx), \]
where
\[ M_i(t, x) = N_i(t, x) \int_0^\infty Y_i(t, u) \lambda(u|Z, W; t) du. \]
By the martingale central limit theorem and a simple evaluation of the predictable variation, the null distribution of \( n^{-1/2} U_0(t) \) is asymptotically zero-mean normal with variance \( \xi_0(t) \). It then follows from the local asymptotic normality of regular parametric models (Bickel, Klaassen, Ritov, and Wellner 1993, pp. 16–17) that under the sequence of contiguous alternatives, \( H_i^{(n)} \),
\[ \beta_i^{(n)} = n^{-1/2} b_i, n^{-1/2} U_0(t) \] is asymptotically normal with mean \( b \xi_0(t) \) and variance \( \xi_0(t) \), which is exactly the limiting distribution of \( n^{-1/2} U(t; \hat{\gamma}_i) \) with \( q(t, x) = q_0(x) \).

**Proof of Theorem 3**

In view of (A.2),
\[ m(t) = \sum_{i=1}^n \int_0^t [q(t, x) \{ Z_i - e_Z(t, x) \} \\
- D'(t) \Omega^{-1}(t) \{ W_i - e_W(t, x) \} ] M_i^{(n)}(t, dx). \] (A.5)
It follows from (A.4) that
\[ n^{-1/2} U(t) = n^{-1/2} m(t) + b \xi(t) + o_p(1). \] (A.6)
Thus the conditional distribution of \( n^{-1/2} U(t) \) or \( S(t) \) will follow directly from that of \( n^{-1/2} m(t) \). Note that (A.5) can be expressed as
\[ m(t) = \sum_{i=1}^n \int_{(t-R_i)^+}^{(t+R_i)^-} [q(t, x) \{ Z_i - e_Z(t, x) \} \\
- D'(t) \Omega^{-1}(t) \{ W_i - e_W(t, x) \} ] dM_i^{(n)}(x), \]
where \( M_i^{(n)}(x) = M_i^{(n)}(\infty, x) \). Because \( M_i^{(n)}(x) \) (\( i = 1, 2, \ldots, n \)) are martingales with respect to the \( \sigma \)filtration generated by the events observed up to \( x, m(t) \) is a martingale with respect to the \( \sigma \)filtration generated by the events observed up to \( t \); that is, \( F_t \). By the martingale central limit theorem, together with a simple calculation of predictable variation, \( n^{-1/2} m(t) \) converges weakly to a zero-mean Gaussian martingale with respect to \( F_t \), and the limiting variance function is \( \xi_x(t) \). Therefore, the conditional distribution of \( n^{-1/2} m(t) \) given \( F_t \) is asymptotically normal with mean \( n^{-1/2} m(t) \) and variance \( \xi_x(t) - \xi_x(t) \). The first part of the theorem then follows from (A.6) on noting that \( \xi_x(t) = \xi(t) \).

When entry time is independent of all other random variables, \( E_Z(t, x, \gamma_0) \) and \( E_W(t, x, \gamma_0) \) converge to \( e_Z(x) \) and \( e_W(x) \).
which are free of \( t \). By the arguments used in the derivation of (A.4),

\[
n^{-1/2} U^2(t; \gamma_0) = n^{-1/2} \sum_{i=1}^{n} \int_{0}^{t} [q(\tau, x)\{Z_i - e_z(x)\} - D'(\tau)\Omega^{-1}(\tau)\{W_i - e_W(x)\}] x M^n(t, dx) + b \xi(t) + o_p(1) \tag{A.7}
\]

Note that \( n^{-1/2}\{U^2(t; \gamma_0) - U(\tau)\} = o_p(1) \), although in general the asymptotic equivalence does not hold for \( t \) that follows from the arguments given in the preceding paragraph that the first term on the right side of (A.7) is an \( \mathcal{F}_t \)-martingale. A direct application of the martingale central limit theorem, together with a simple calculation of predictable variation, shows that \( n^{-1/2} U^2(t; \gamma_0) \) converges weakly to a Gaussian \( \mathcal{F}_t \)-martingale with mean function \( b \xi(t) \) and variance function given in (10). The second part of the theorem then follows immediately from the martingale feature of \( U^2(t; \gamma_0) \).

[Received August 1997. Revised August 1998.]

REFERENCES


