Statistics and Its Interface Volume 5 (2012) 415–423

Futility stopping in clinical trials

Pei He, Tze Leung Lai* and Olivia Y. Liao

Early stopping due to futility, also referred to as a go/no-go decision, during interim analysis has become an important feature of clinical trial designs. Current methods for futility stopping in literature are mostly based on conditional power or predictive power in conjunction with the theory of stochastic curtailment or group sequential design. They have certain drawbacks that have been noted in literature. Herein we describe a new approach to futility stopping in clinical trial designs and the statistical theory underlying this approach. Simulation studies and theoretical analysis show the advantages of the approach in both parametric and nonparametric problems.

AMS 2000 subject classifications: Primary 62L05; secondary 62F03.
Keywords and phrases: Conditional power, Efficient, Group sequential testing, Survival endpoint.

1. INTRODUCTION

Since the late 1970s, interim analyses of the accumulating data in a long-term clinical trial have become increasingly popular and are now standard in clinical trial designs to compare a new treatment with a control. While there have been many important advances in group sequential tests with prescribed type I error probability of falsely rejecting the null hypothesis that treatment is not better than control, theoretical developments in futility stopping (as opposed to early stopping for efficacy that may inflate the type I error) seem to have lagged behind. The seminal paper of Lan, Simon and Halperin [16] introduced the conditional power approach and provided a fundamental bound on the loss of power at a given alternative due to futility stopping by considering the conditional power at that alternative. The need to choose an alternative in the condtional power approach was subsequently circumvented by using the predictive power approach or by using maximum likelihood or other methods to estimate the actual parameter at interim analysis so that the estimate can be used to substitute for the alternative in the conditional power approach. Section 2 gives a review of the conditional power and predictive power approaches to futility stopping and certain issues with both approaches.

*Corresponding author.

Section 3 develops a statistical theory for futility stopping, which relates futility to evidence against the hypothesis that treatment is better than control by at least some prescribed amount. For testing the one-sided null hypothesis $H_0 : u(\theta) \le u_0$ in a multiparameter exponential family, this theory formulates futility stopping via rejection of the hypothesis $u(\theta) \ge u_1$, in which u_1 is an "implied alternative," that is, the alternative at which the reference test of H_0, with type I error α and taking no more than M observations, has power $1 - \hat{\alpha}$. Using this formulation, we describe in Section 3 a new methodology for futility stopping, first in the exponential family setting and then in nonparametric and semiparametric models. Further complications for time-sequential trials with survival endpoints, which have been pointed out in [15, 18], are also addressed by an extension of the basic approach. Simulation studies are given in Section 4 and show the advantages of the proposed approach to futility stopping. Further discussion and concluding remarks are given in Section 5.

2. CONDITIONAL AND PREDICTIVE POWER APPROACHES

The motivation underlying conditional and predictive power is to forecast the outcome of a given test, called a *reference test*, of a statistical hypothesis H_0 from the data D_t up to the time t when such a prediction is made. Since the outcome is binary (i.e. whether to reject H_0 or not), the forecast can be presented as the probability of rejecting H_0 at the end of the study given D_t. However, this probability has to be evaluated under some probability measure. In the context of hypothesis testing in a parametric family $\{P_\theta, \theta \in \Theta\}$, [16] proposed to consider the conditional power

$$p_t(\theta) = P_\theta(\text{Reject } H_0 | D_t). \tag{1}$$

Subsequently, [5] and [21] found it more appealing to put a prior distribution on θ and consider the posterior probability of rejecting H_0 at the end of the trial given D_t, and therefore advocated to consider the predictive power

$$P_t = P(\text{Reject } H_0 | D_t) = \int p_t(\theta) d\pi(\theta | D_t) \tag{2}$$

where $\pi(\theta|D_t)$ is the posterior distribution of θ. This idea had been proposed earlier by Herson [9].

For the problem of testing the one-sided hypothesis $H_0 : \theta \le \theta_0$, the predictive power approach to futility stopping

terminates the study if $P_t \le \gamma$, for some threshold $0 < \gamma < \frac{1}{2}$, at time t of interim analysis. Similarly the conditional power approach chooses an alternative $\theta_1 > \theta_0$ and stops the study if $p_t(\theta_1) \le \gamma$. It is shown in [16] that if the reference test, which may be sequential, has power $1-\beta$ at θ_1, then adding such futility stopping feature to the test at all times of interim analysis still has power $\ge (1-\beta)/(1-\gamma)$; see also [12, p. 206–207].

While the conditional power approach to futility stopping requires specification of an alternative θ_1, the predictive power approach requires specification of a prior distribution π. It is often difficult to come up with such a specification in practice. On the other hand, one can use D_t to estimate the actual θ by maximum likelihood or other methods, as suggested by Lan and Wittes [17]. For normal observations X_i with common unknown mean θ and known variance σ^2, using Lebesgue measure on the real line as the improper prior for θ yields the sample mean $\bar{X}_t$, as the posterior mean and also the MLE. In this case, for the fixed sample size test that reject $H_0 : \theta = 0$ if $\sqrt{n}\bar{X}_n \ge \sigma z_\alpha$, this predictive power is

$$\Phi\left(\sqrt{\frac{t}{n-t}}\left(\frac{\sqrt{n}}{\sigma}\bar{X}_t - z_\alpha\right)\right), \tag{3}$$

and the conditional power is

$$p_t(\bar{X}_t) = \Phi\left(\sqrt{\frac{n}{n-t}}\left(\frac{\sqrt{n}}{\sigma}\bar{X}_t - z_\alpha\right)\right); \tag{4}$$

see [12, p. 211]. Here and in the sequel, we use Φ to denote the standard normal distribution function and $z_\alpha = \Phi^{-1}(1-\alpha)$.

Although using the conditional or predictive power to guide early stopping for futility is intuitively appealing, there is no statistical theory for such choice of the stopping criterion. In fact, using the MLE as the alternative already pre-supposes that the MLE falls outside the null hypothesis, and a widely used default option is to stop when the MLE belongs to H_0, which is consistent with (4) that falls below the type I error α in this case. However, this ignores the uncertainty in the estimate and can lose substantial power due to premature stopping, as shown in the simulation studies of [2, 3] on adaptive designs that use this kind of futility stopping. Pepe and Anderson [20] have proposed to adjust for this uncertainty by using $\bar{X}_t + \sigma/\sqrt{t}$ instead of $\bar{X}_t$ to substitute for θ_1 in the conditional power approach.

Instead of estimating the alternative during interim analysis, one can focus on a particular alternative θ_1 and consider the conditional power $p_t(\theta_1)$ or the predictive power with a prior distribution concentrated around θ_1. Although [16] has shown that adding futility stopping to the reference test of $H_0 : \theta \le \theta_0$ if $p_t(\theta_1) \le \gamma$ does not decrease the power of the reference test at θ_1, by more than a factor of $\gamma/(1-\gamma)$, there is no statistical theory justifying why one should use a conditional instead of an unconditional test of $\theta \ge \theta_1$. Furthermore, as noted earlier, this approach leaves open the problem of how θ_1 should be chosen for stopping a study due to futility.

3. FUTILITY STOPPING THEORY

In this section we develop a statistical theory of early stopping for futility. To fix the ideas, consider a confirmatory trial sponsored by a pharmaceutical company to demonstrate the efficacy of a new drug for its approval. The null hypothesis assumed by the regulatory agency is $H_0 : u(\theta) \le u_0$, where θ is a parameter vector and u is a smooth function of θ so that H_0 represents that the treatment is not efficacious. For example, $\theta = (p_1, p_2)$ in the case of binary responses, where p_1 is the response probability of the new treatment and p_2 is that of the standard treatment, and $u(\theta) = p_1 - p_2$, $u_0 = 0$. This formulation also includes nonparametric tests by allowing θ to be infinite-dimensional, e.g., $\theta = (F_1, F_2)$ with F_i being the distribution function of the new ($i = 1$) or standard ($i = 2$) treatment. Here smoothness of u means that it is compactly differentiable; see [6]. In Sections 3.1 and 3.2, we focus on the parametric case involving exponential families for which we can apply results from the theory of [14] on efficient group sequential tests. Section 3.3 provides extensions to nonparametric and semiparametric tests, and Section 3.4 addresses additional issues for time-sequential trials with survival endpoints.

3.1 Alternative implied by the maximum sample size constraint

Because of time and resource constraints, the sample size of a clinical trial cannot exceed some prescribed upper bound M. Lai and Shih [14] introduced the concept of an "implied alternative," which is the alternative implied by this constraint, for testing $H_0 : \theta \le \theta_0$ based on i.i.d. observations $X_1, X_2, \ldots$ from a one-paramter exponential family with density $f_\theta(x) = e^{\theta x - \psi(\theta)}$. In this setting, application of the Neyman-Pearson lemma yields the fixed sample size test that rejects H_0 if $S_M \ge c_\alpha$ as the uniformly most powerful (UMP) level-α test for every alternative $\theta > \theta_0$, where $S_M = \sum_{i=1}^M X_i$ and c_α is chosen such that $P_{\theta_0}(S_M \ge c_\alpha) = \alpha$. In particular, it has maximal power, among all level-α tests (including sequential ones) that take no more than M observations, at the implied alternative $\theta(M)$ where the above level-α test has prescribed power $1-\tilde{\alpha}$. In fact, the sample size M can be determined by $P_{\theta_1}(S_M \ge c_\alpha)$ when $\theta_1 = \theta(M)$ is given.

The successful outcome of the trial, from the pharmaceutical company's viewpoint, is rejection of the one-sided hypothesis H_0 by using a test that maintains the specified type I error. Although stopping early for futility would not increase the type I error, it would decrease the power of the test. In paticular, adding futility stopping to the test that rejects H_0 if $S_M \ge c_\alpha$ would lose the UMP property. Since the UMP test has the desired power $1-\tilde{\alpha}$ at $\theta(M)$, we would

like the power to be still around $1-\tilde{\alpha}$ at $\theta(M)$ when such futility stopping is introduced. Thus, in this one-parameter exponential family setting, one can study futility stopping via the problem of testing $H_0 : \theta \le \theta_0$ versus $H_1 : \theta \ge \theta(M)$, with type I error α and type II error slightly more than $\tilde{\alpha}$ and taking no more than M observations.

3.2 Group sequential testing theory

For the problem of testing $H_0 : \theta \le \theta_0$ versus $H_1 : \theta \ge \theta_1$ in the one-parameter exponential family, Lai and Shih [14] start by extending Hoeffding's lower bound [10] for the expected sample size $E_\theta(T)$ with error probabilities α at θ_0 and $\tilde{\alpha}$ at θ_1 to the group sequential setting, in which stopping can only occur at times of interim analysis with sample sizes $n_1 < \cdots < n_k = M$. They show that this lower bound can be attained asymptotically as the error probabilities approach 0 by a group sequential version of Lorden's 2-SPRT [19], which runs simultaneously the sequential probability ratio test (SPRT) of the simple null θ_0 versus the simple alternative θ and the SPRT of θ_1 versus θ, and which stops as soon as one of the SPRTs rejects the corresponding null hypothesis. They then modify this test to allow sequential updating of θ by maximum likelihood during the course of the trial. The likelihood ratio of θ to θ_j is replaced by the generalized likelihood ratio (GLR) of $\hat{\theta}_{n_i}$ to θ_j, $(j = 0, 1)$ at the ith interim analysis. In particular, the test stops early at the ith interim analysis $(1 \le i \le k-1)$ and rejects $H_1 : \theta \ge \theta_1$ if

$$\hat{\theta}_{n_i} < \theta_1 \quad \text{and} \quad n_i I(\hat{\theta}_{n_i}, \theta_1) \ge \tilde{b}, \tag{5}$$

where $\tilde{b}$ is chosen such that

$$P_{\theta_1}\{(5) \text{ holds for some } 1 \le i \le k-1\} = \epsilon\tilde{\alpha} \tag{6}$$

for some $0 < \epsilon < \frac{1}{2}$, in which $I(\theta, \lambda)$ is the Kullback-Leibler information number $(\theta-\lambda)\psi'(\theta) - \{\psi(\theta) - \psi(\lambda)\}$ and therefore $n_i I(\hat{\theta}_{n_i}, \theta_1)$ is the logarithm of the GLR of $\hat{\theta}_i$ to θ_1. Note that (5) can be regarded as testing θ_1 against the MLE $\hat{\theta}_{n_i} < \theta_1$ and that rejection of H_1 is the same as acceptance of H_0, so the sequential GLR test of H_1 is used for futility stopping in testing H_0.

Early stopping for efficacy, with H_0 rejected, occurs if

$$\hat{\theta}_{n_i} > \theta_0 \quad \text{and} \quad n_i I(\hat{\theta}_{n_i}, \theta_1) \ge b. \tag{7}$$

In case stopping does not occur in the first $k-1$ analyses, reject H_0 if $S_{n_k} \ge c$, where b and c are so chosen that the test has error probability α at θ_0 of falsely rejecting H_0. Note that $S_{n_k} \ge c$ can be written in the form (7) with $i = k$ and b replaced by $\tilde{c}$, and that (7) can be regarded as testing θ_0 against the MLE $\hat{\theta}_{n_i} > \theta_0$. With $\theta_1 = \theta(M)$, Theorem 3 of [14] shows that the test, which is called the *modified Haybittle-Peto test*, has asympototically minimal expected sample size $E_\theta(T)$ at the true parameter θ and has power $1-\tilde{\alpha}-\kappa_\epsilon\tilde{\alpha}+o(\tilde{\alpha})$ at θ_1 as $\alpha+\tilde{\alpha} \to 0$, where κ_ϵ is a constant depending on ϵ.

Section 3.4 of [14] has extended this group sequential testing theory to the multiparameter exponential family $f_\theta(x) = e^{\theta^T x - \psi(\theta)}$ and to multi-arm settings. The null hypothesis has the more general form $H_0 : u(\theta) \le u_0$. The Kullback-leibler information number becomes $(\theta-\lambda)^T\nabla\psi(\theta) - \{\psi(\theta)-\psi(\lambda)\}$, where ∇ is the gradient vector. Suppose $I(\theta, \lambda)$ is increasing in $|u(\lambda) - u(\theta)|$ for every fixed θ. Then we can still define the alternative u_1 implied by the maximum sample size M and the type II error probability $\tilde{\alpha}$ of the reference test. Let $u_1 = u_1(M)$ be such tht the fixed sample size GLR test of H_0 with type I error probability α and sample size M has power

$$\inf_{\theta : u(\theta) = u_1} P_\theta(\text{GLR test rejects } H_0) = 1 - \tilde{\alpha}; \tag{8}$$

see [3, Section 2.1]. Therefore futility stopping can again be carried out as before, using the sequential GLR test of $H_1 : u(\theta) \ge u_1$. In multi-arm clinical trials, for which different numbers of patients are assigned to I different treatments, the GLR statistic at the jth interim analysis is e^{Λ_j}, where

$$\Lambda_j = \sum_{i=1}^{I} n_{ij}\{\hat{\theta}_{i,n_{ij}}^T \bar{X}_{i,n_{ij}} - \psi(\hat{\theta}_{i,n_{ij}})\} - \sup_{u(\theta_1,\ldots,\theta_I) = u_0} \sum_{i=1}^{I} n_{ij}\{\theta_i^T \bar{X}_{i,n_{ij}} - \psi(\theta_i)\},$$

in which n_{ij} is the total number of observations from the ith population up to the time of the jth interim analysis. Let $n_j = n_{1j} + \cdots + n_{Ij}$. The asymptotic theory for the modified Haybittle test is extended in Section 3.4 of [14] to the case that uses adaptive randomization such that $n_{ij} = p_i n_j + O_p(\sqrt{n_j})$ under $u(\theta) = u_0$ or under $u(\theta) = u_1$, in which $p_1, \ldots, p_I$ are nonnegative constants that sum up to 1 and can differ for the cases u_0 and u_1. Zhu and Hu [23] have recently demonstrated the advantages of using adaptive randomization over traditional equal randomization in group sequential trials.

3.3 Extensions to nonparametric and semiparametric tests

In many confirmatory clinical trials, M is large and the treatment effect is assumed to belong to the framework of "local alternatives" for sample size calculation that justifies the choice of M. These local alternatives lead to "locally asymptotically normal" (LAN) families for the sampling distributions of the parametric, or nonparametric, or semiparametric test statistics used; see [1]. Because of its proximity to H_0, the MLE of a local alternative has substaintial probability of falling in H_0, and therefore one can lose considerable power by replacing a local alternative by its MLE in the conditional power $p_t(\theta)$, as pointed out in Section 2.

We first consider nonparametric group sequential tests of $H_0 : u(F, G) \le 0$, where F is the distribution function of the outcome of a new treatment and G is that of the standard treatment (or placebo), and $u(F, F) = 0$. Let n_i' be the sample size of the new treatment and n_i'' be that of the standard treatment at the ith interm analysis so that $n_i = n_i' + n_i''$, and let $X_1, \ldots, X_{n_i'}$ and $Y_1, \ldots, Y_{n_i''}$ be the corresponding outcomes. Let $\hat{F}_{n_i'}$ be the empirical distribution function of $X_1, \ldots, X_{n_i'}$, and $\hat{G}_{n_i''}$ be that of $Y_1, \ldots, Y_{n_i''}$. As shown in [13], commonly used two-sample nonparametric test statistics can be written in the form of a *generalized Chernoff-Savage statistic*

$$T_i = \int_{-\infty}^{\infty} J_i(\hat{F}_{n_i'}(x), \hat{G}_{n_i''}(x)) d\hat{F}_{n_i'}(x), \tag{9}$$

where $J_i : \{0, 1/n_i', 2/n_i' \ldots 1\} \times \{0, 1/n_i'', 2/n_i'' \ldots 1\} \to \mathbb{R}$ satisfies

$$\frac{1}{n_i'} \sum_{l=1}^{n_i'} \sup_{y \in \{1/n_i'', \ldots, 1\}} \left| J_i\left(\frac{1}{n_i'}, y\right) - J\left(\frac{1}{n}, y\right) \right| \to 0 \tag{10}$$

as $n_i' \to \infty$, and $J : [0, 1] \times [0, 1] \to \mathbb{R}$ is twice continuously differentiable except possibly at $(0, 0)$ and $(1, 1)$ and satisfies certain regularity conditions near $(0, 0)$ and $(1, 1)$. In this case, the function $u(F, G)$ in $H_0 : u(F, G) \le 0$ is given by

$$u(F, G) = \int_{-\infty}^{\infty} J(F(x), G(x)) dF(x). \tag{11}$$

Since subjects are randomly assigned to the new or standard treatment,

$$n_i'/n_i \xrightarrow{p} 1/2, \text{ i.e., } n_i'' = n_i'(1 + o_p(1)). \tag{12}$$

Under (12), T_i has the representation

$$\begin{aligned} T_i = u(F, G) &+ \frac{1}{n_i'} \sum_{l=1}^{n_i'} (\psi(X_i) - E\psi(X_i)) \\ &+ \frac{1}{n_i''} \sum_{l=1}^{n_i''} (\psi^*(Y_i) - E\psi^*(Y_i)) + R_i \end{aligned} \tag{13}$$

where $R_i = o_p(1/\sqrt{n_i'})$ and

$$\begin{aligned} \psi(x) &= J(F(x), G(x)) - \int_0^x \frac{\partial J}{\partial x}(F(t), G(t)) dF(t), \\ \psi^*(y) &= -\int_0^y \frac{\partial J}{\partial y}(F(t), G(t)) dF(t); \end{aligned} \tag{14}$$

see [4] and [13], which sharpens the proof in [4], to obtain a stronger result on R_i. For the present application, the convergence result $R_i = o_p(1/\sqrt{n_i'})$ suffices. Therefore, under equal randomization to the two treatments, $n_i' T_i$ behaves asymptotically like a normal random walk under H_0 and under local alternatives, and the problem of testing $H_0 : u(F, G) \le 0$ versus $H_1 : u(F, G) \ge \delta$ can be approximated by that of group sequential testing of $H_0' : \mu \le 0$ versus $H_1' : \mu \ge \delta$ based on i.i.d. normal random variables $Z_1, Z_2, \ldots$ with mean $\mu = u(F, G)$ and variance

$$\sigma^2 = \mathrm{Var}_{F=G}(\psi(X)) + \mathrm{Var}_{F=G}(\psi^*(Y)), \tag{15}$$

where ψ and ψ^* are given by (14). Note that $F = G$ is the boundary case of H_0 and that $F(X)$ and $G(Y)$ are Uniform$(0, 1)$ random variables. Under local alternatives, the asymptotic variance of T_i is the same as that under $F = G$, and therefore the variance formula (15) still holds for local alternatives. Further details and examples are given in Section 4.2.

The limiting normal random walk model and the more general Gaussian process model with independent increments have been derived by [11] and [12, Chapter 11] for efficient score statistics in parametric and Cox regression models. We have shown above that the limiting joint distribution holds generally for nonparametric test statistics, which may not be efficient score statistics, under the null hypothesis and local alternatives. Similarly, in the normal linear models considered by Jennison and Turnbull [11, Section 3], we can replace the normal models for the random errors by nonparametric models, leading to semiparametric regression models. When the regression parameters are estimated by least squares in this case, the score statistics or Wald statistics used for testing the regression parameters are not efficient, but the asymptotic joint normality of the test statistics at interim analysis still holds for the local alternatives. Therefore the theory for futility stopping developed in the preceding section for the exponential family (which includes the normal family) can be extended to the nonparametric or semiparametric setting. This is analogous to how group sequential tests developed for the prototypical normal case are extended to more general settings in [12].

3.4 Futility stopping in time-sequential trials with survival outcomes

Scharfstein, Tsiatis and Robins [**?**] have extended the preceding results under the null hypothesis and local alternatives to efficient time-sequential score tests in semiparametric models. In particular, in the Cox regression model with regression parameter β, the efficient score statistic $S_n(t)$, which is the first derivative of the partial likelihood ratio statistic with respect to β, is locally asymptotically normal. Here n is the sample size and t is the time of an interim analysis. Specifically, $n^{-1/2} S_n(t)$ converges in distribution, as $n \to \infty$, to a Gaussian process with independent increments with variance $V(t)$ under the null hypothesis $\beta = 0$ and local alternatives. The mean of the limiting Gaussian process is 0 under $\beta = 0$ and is $\delta V(t)$ under local alternatives in the proportional hazards (or Cox) regression model. Therefore, the null variance of $S_n(t_i)$, which is

approximately $nV(t_i)$ at the ith interim analysis takes the place of the sample size n_i in the preceding discussion.

The power calculations at the design stage of a time-sequential trial with survival endpoint typically assume a working model of survival functions $\bar{F} = 1 - F$ and $\bar{G} = 1 - G$, the accrual pattern and the censoring rates per year. The working model embeds the null case $\bar{F} = \bar{G}$ in a semiparametric family whose parameters are fully specified for the alternative hypothesis, under which the study duration and sample size of the two-sample semiparametric test are shown to have some prescribed power. Illustrative examples are given in Section 4.3. The two-sample test statistic $S_n(t)$ is usually chosen to be an efficient score statistic or its asymptotic equivalent in the working model. Unlike the nonparametric two-sample test of $H_0 : u(F, G) \le 0$ considered above, in which the asymptotic variance of $n_i' T_i$ is proportional to the sample size $n_i' \approx n_i/2$ at the ith interim analysis in view of (13), the asymptotic null variance $nV(t_i)$ of $S_n(t_i)$ depends not only on the survival distribution but also on the accrual rate and the censoring distribution up to the time t_i of the ith interim analysis. The observed patterns, however, may differ substantially from those assumed in the working model for the power calculations at the design stage. In addition, the working model under which the test statistic is semiparametrically efficient (e.g., the proportional hazards model when a logrank test is used) may not actually hold. In this case, as the sample size n approaches ∞, the limiting distribution of $\sqrt{n}S_n(t)$ is still normal with mean 0 and variance $V(t)$ under $F = G$ and has independent increments, but under local alternatives, the mean $\mu(t)$ of the limiting normal distribution of $\sqrt{n}S_n(t)$ may not be linear in $V(t)$, and may level off or even decrease with increasing $V(t)$; see [7].

For the futility stopping decision at interim analysis, we can consider local alternatives, which suggest using the test $H_0 : \mu(t_i) \le 0$ for $1 \le i \le k$ versus $H_\delta : \mu(t_i) \ge \delta V(t_i)$ for some i, as discussed in the first paragraph of this section. We choose the same δ as that used in the design stage to determine the sample size and trial duration, since we do not want to have substantial power loss at or near the alternative assumed at the design stage. Even when the working model does not actually hold, for which $\mu(t)/V(t)$ may vary with t, using it to determine the implied alternative for futility stopping only makes it more conservative to stop for futility because $\mu(t)$ tends to level off or even decrease instead of increasing linearly with $V(t)$. It remains to consider how to update, at the ith interim analysis, the estimated value of the "maximum information" $nV(t^*)$ (and also $nV(t_j)$ for $j > i$ if the reference test is time-sequential) after observing accrual, censoring and survival patterns that differ substantially from those assumed at the design stage. Our strategy is to replace $V(t)$ by the estimated $\hat{V}(t)$ for $t > t_i$ in the efficient score test of H_δ that involves these values, but not to estimate $\mu(t)$ for $t > t_i$ because we are in the setting of local alternatives with small δ (of the order $1/\sqrt{n}$).

Bayesian modeling provides a natural updating scheme for estimating at time t_i of interim analysis based on observations up to t_i, the null variance $V_n(t)$ of the score statistic $S_n(t)$ for $t > t_i$. Following [22], we use Dirichlet process priors for the distribution function $(F + G)/2$ and for the censoring (i.e., patient withdrawal or loss in follow-up) distribution. Note that the null variance $V_n(t)$ is generated by the accrual rate, the censoring distribution, and the survival distributions F and G that are assumed to be equal. The parameter α, which is a finite measure on $\mathbb{R}_+ = (0, \infty)$, of the Dirichlet process prior for $1 - H$, where $H = (\bar{F} + \bar{G})/2$, can be chosen to be some constant times the assumed parametric model, that is used for power calculation at the design stage, where the constant is $\alpha(\mathbb{R}_+)$ that reflects the strength of this prior measure relative to the sample data. At the ith interim analysis, let n_i be the total number of subjects who have been accrued and let

$$Z_j^{(i)} = \min(T_j, \xi_j, t_i - \tau_j), \qquad \delta_j^{(i)} = I_{\{Z_j^{(i)} = T_j\}},$$

$j = 1, \ldots, n_i$, where T_j is the actual survival time of the jth patient, τ_j is the patient's entry time and ξ_j is the censoring time. By re-arranging the observations, we can assume without loss of generality that $Z_1^{(i)}, \ldots, Z_k^{(i)}$ are the uncensored observation, and let $Z_{[k+1]}^{(i)} < \cdots < Z_{[m]}^{(i)}$ denote the distinct ordered censored observations. Let

$$N_i(u) = \sum_{j=1}^{n_i} I_{\{Z_j^{(i)} \ge u\}}, \qquad N_i^+(u) = \sum_{j=1}^{n_i} I_{\{Z_j^{(i)} > u\}},$$

$$\lambda_i(u) = \sum_{j=1}^{n_i} I_{\{Z_j^{(i)} = u, \delta_j = 0\}}, \qquad Z_{[k]}^{(i)} = 0, \qquad Z_{[m+1]}^{(i)} = \infty.$$

As shown in [22], for $Z_{[l]}^{(i)} \le u < Z_{[l+1]}^{(i)}$, the Bayes estimate of $H(u)$ at the ith interim analysis is given by

(16)

$$\hat{H}_i(u) = \frac{\alpha(u, \infty) + N_i^+(u)}{\alpha(\mathbb{R}_+) + n_i} \times \prod_{j=k+1}^{l} \left\{ \frac{\alpha[Z_{[j]}^{(i)}, \infty) + N_i(Z_{[j]}^{(i)})}{\alpha[Z_{[j]}^{(i)}, \infty) + N_i(Z_{[j]}^{(i)}) - \lambda_i(Z_{[j]}^{(i)})} \right\}.$$

Similarly, for updating the estimate $\hat{C}$ of the censoring distribution, we can interchange the roles of T_j and ξ_j above and replace α by α_c that is associated with the specification of the censoring distribution at the design stage. The accrual rates for the period prior to t_i have been observed and those for the future years can use what is assumed at the design stage. Since $V_n(t) = V_n(t_i) + [V_n(t) - V_n(t_i)]$, we can estimate $V_n(t)$ by $V_n(t_i) + \mathrm{E}[V_n^*(t) - V_n^*(t_i) | \hat{H}, \hat{C}]$, in which the expectation E assumes the updated accrual rates and can be computed by Monte Carlo simulations to generate the observations (Z_j^*, δ_j^*) that are independent of the

Table 1. Comparative study in normal case

θ	FSS	CP(MLE)		CP(MLE+se)		PP		GLR_f		$GLR_{e,f}$	
	Power	Power	E(#)	Power	E(#)	Power	E(#)	Power	E(#)	Power	E(#)
0	0.05	0.03	1.60	0.04	2.32	0.03	1.79	0.05	3.15	0.05	3.10
$\theta_1 = 0.13$	0.80	0.57	3.58	0.72	4.41	0.62	3.88	0.79	4.80	0.78	3.74
$1.2\theta_1$	0.91	0.69	3.95	0.84	4.64	0.75	4.21	0.90	4.90	0.89	3.42
$2\theta_1$	1.00	0.95	4.81	0.99	4.97	0.97	4.89	1.00	5.00	1.00	1.97

Table 2. Comparative study in nonparametric case

θ	FSS	CP(MLE)		CP(MLE+se)		PP		GLR_f		$GLR_{e,f}$	
	Power	Power	E(#)	Power	E(#)	Power	E(#)	Power	E(#)	Power	E(#)
0	0.05	0.03	1.59	0.04	2.31	0.03	1.66	0.05	3.16	0.05	3.11
$\theta_1 = 0.13$	0.77	0.53	3.49	0.68	4.32	0.55	3.58	0.75	4.77	0.74	3.83
$1.2\theta_1$	0.88	0.66	3.86	0.82	4.59	0.68	3.94	0.87	4.88	0.86	3.56
$2\theta_1$	1.00	0.92	4.70	0.98	4.95	0.93	4.72	0.99	4.99	1.00	2.20

Table 3. Time-sequential example

	Hazard rate of F	FSS	GLR_f		$GLR_{e,f}$	
		Power	Power	E(#)	Power	E(#)
(1)	$\lambda_0 = 1/3$ $(F = G)$	0.05	0.05	3.35	0.05	3.31
(2)	$\lambda_0/1.4$	0.82	0.81	4.90	0.80	3.72
(3)	$\lambda_0/1.5$	0.92	0.91	4.94	0.91	3.37
(4)	$\lambda_0/1.65$	0.98	0.98	4.98	0.98	2.85
(5)	$\lambda_0/4$ for $0 \le s \le 1$, λ_0 for $s > 1$	0.93	0.93	5.00	0.97	1.55
(6)	$\lambda_0/5$ for $0 \le s \le 1$, λ_0 for $s > 1$	0.96	0.96	5.00	0.99	1.35
(7)	$\lambda_0/4.5$ for $s \le 1$ or $s \ge 6$, $\lambda_0/0.9$ for $1 < s < 6$	0.84	0.84	5.00	0.95	1.55

$(Z_j^{(i)}, \delta_j^{(i)})$ observed up to time t_i. Note that we can use the limiting independent increments property of $V_n(t)/n$ to simplify these computations and those for the stopping boundaries.

4. SIMULATION STUDIES

This section describes simulation studies of the futility stopping approach developed in Section 3. The results are given in Tables 1, 2 and 3; each result is based on 10,000 simulations. The reference test has a fixed sample size. We also consider in the last column of each table the case where the test can stop early for either efficacy or futility. Tables 1 and 2 compare the proposed methods in Sections 3.2 and 3.3 with the conditional power methods of [17] and [20]. These conditional power methods are denoted by CP(MLE) and CP(MLE + se), respectively, and use the stopping criterion $\mathrm{CP}(\cdot) \le \gamma = 0.3$ suggested by [20]. For comparison with the predictive power (PP) method described in Section 2, we replace conditional power by the predictive power associated with the flat prior. Table 3 considers the time-sequential setting in Section 3.4, for which the reference test has fixed sample size and study duration.

4.1 A comparative study for the prototypical normal mean case

This simulation study considers the prototypical normal case in which the outcomes of the treatment arm and the placebo arm follow $N(\mu_X, 1)$ and $N(\mu_Y, 1)$, respectively. The problem is to test $H_0 : \theta = \mu_X - \mu_Y \le 0$, with type I error $\alpha = 0.05$ and power $1 - \tilde{\alpha} = 0.8$ when $\mu_Y = 1.5$. The trial involves $M = 1,000$ subjects randomized to both arms with probability 1/2 each, and interim analyses are planned with $n_i = 200, 400, 600$, and 800 subjects, for $i = 1, 2, 3, 4$. The fixed sample size test has power 0.8 at the alternative 0.13, which is the implied alternative $\theta(M)$ and will be denoted simply by θ_1. We apply the group sequential GLR test described in the first paragraph of Section 3.2 with $\epsilon = 1/3$ to perform futility stopping. This test will be denoted by GLR_f, in which the subscript f stands for early stopping for futility. Similarly, $GLR_{e,f}$ denotes the group sequential GLR test involving both efficacy and futility stopping. For $GLR_{e,f}$ we use a non-binding futility boundary that does not consider the possibility of futility stopping in determining the efficacy boundary. Table 1 gives the power and the expected number of groups $E(\#)$ for GLR_f or $GLR_{e,f}$.

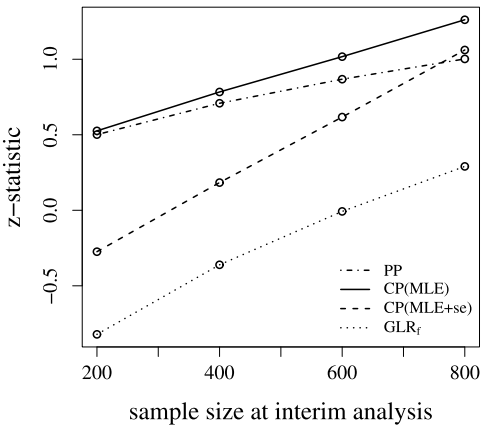

Figure 1. Futility stopping thresholds.

Since the group size is 200, the expected sample size is $200E(\#)$.

The first row ($\theta = 0$) of Table 1 shows that applying futility stopping reduces $E(\#)$ from 5 to 1.6−3.15. The other rows show the performance of each test at the implied alternative θ_1 and at the larger alternatives $1.2\theta_1$ and $2\theta_1$. Compared to the fixed sample size test, the group sequential GLR tests only lose 1%−2% power, while the conditional power and predictive power tests lose much more power, although they have smaller $E(\#)$. This fact is also illustrated in Fig. 1. The stopping criterion of the conditional or predictive power test and the group sequential GLR$_f$ test can be transformed into thresholds for the standardized z-statistics

$$\frac{\bar{X}_{n_i/2} - \bar{Y}_{n_i/2}}{\sqrt{4/n_i}} \le Z_{\mathcal{F}}(n_i),$$

assuming approximately equal assignments of the n_i subjects to the two treatments. In Fig. 1, $Z_{\mathcal{F}}(n_i)$ is plotted against sample size n_i for $\mathcal{F}$ = PP, CP(MLE), CP(MLE + se), GLR$_f$. It shows that $Z_{\mathcal{F}}(n_i)$ is markedly smaller for $\mathcal{F}$ = GLR$_f$ than the other methods and that CP(MLE + se) has the second smallest value. Moreover, it is possbile to lower the threshold for CP(MLE + se) by using a smaller value of the threshold γ to trigger futility stopping when CP($\cdot$) falls below the threshold. In fact, replacing $\gamma = 0.3$ by $\gamma = 0.05$ brings the performance of CP(MLE + se) close to that of GLR$_f$ in terms of power and $E(\#)$. Moreover, the steeper slope of $Z_{\text{CP(MLE+se)}}(n_i)$ reflects the adjustment for the uncertainty of the MLE, and such adjustment results in almost 20% more power than CP(MLE) in some cases.

4.2 Futility stopping for two-sample Wilcoxon tests

The second simulation study uses the fixed sample size Wilcoxon test of $H_0: F = G$ as the reference test. In view of (13), we can approximate the two-sample Wilcoxon statistic by a sum of normal random variables $Z_1, Z_2, \ldots$ with mean $P(X < Y)$ and variance 1/12 under H_0 and local alternatives. The simulation study uses the same design, sample size, and parameters as those described in Section 4.1, and assumes the distribution F to be χ^2 with 2 degrees of freedom, and G to be a location shift of F, i.e., $G(x) = F(x-\theta)$. The implied alternative $\theta = 0.182$ is the distance of the location shift, at which the fixed sample size test has power 80% using the aforementioned normal approximation. Table 2 shows results similar to those in Table 1, and in particular, that the conditional and predictive power tests tend to stop earlier for futility at the expense of substantial power loss when compared to the fixed sample size test.

4.3 Futility stopping in time-sequential trials with survival outcomes

The third simulation study is a continuation, with some modifications, of an example in [7]. The simulation study involves $n = 450$ patients who arrive independently and uniformly over a 3-year interval and are randomized to the treatment and placebo arms. There are $k = 5$ analyses at 1.5, 2.5, 3.5, 4.5 years and at 5.5 years when the trial is scheduled to end. Hence the expected study duration is the expected number of stages $E[\#]$ plus 0.5. Censoring due to loss of follow-up is assumed to be exponentially distributed with rate 1/6. The failure-time distribution G for the placebo arm is assumed to be exponential with hazard rate $\lambda_0 = 1/3$, so the median survival is of 3 years. Simulations are conducted under the null hypothesis $F = G$ (case 1) and under proportional hazards alternatives in which F is exponential with hazard rate $\lambda = \lambda_0/1.4,\ \lambda_0/1.5,\ \lambda_0/1.65$ (cases 2–4). Besides proportional hazards alternatives, three other stochastically ordered alternatives are also considered and listed in Table 3 as cases 5–7. In cases 5 and 6, the hazard rate of F is lower than that of G in the first year. In case 7, the hazard rate of G only exceeds that of F in years one to six.

Futility stopping via the sequential GLR method described in Section 3.4 saves almost half of the time in case 1, while losing less than 1 percent of the power compared to the fixed sample size and fixed duration test. For proportional hazards alternative, GLR$_{e,f}$ has power similar to that of GLR$_f$ but saves 1–2 stages (years), since the trial is allowed to stop early for efficacy. It even improves the power and reduces substantial study duration by 3.4–3.6 years in cases 5–7. This phenomenon has been explained by [7]; the expected value of $S_n(t)$ may actually decrease with increasing t, therefore allowing GLR$_{e,f}$ to achieve both savings in time and increase in power over fixed-duration tests.

As pointed out in Section 3.4, our approach to futility stopping via GLR$_f$ or GLR$_{e,f}$ requires updating the estimate of $V(t^*)$, with $t^* = 5.5$, at each interim analysis. We assume exponential densities $2,000e^{-0.6t}$ and $2,000e^{-0.33t}$ for the Dirichlet process priors of $(F+G)/2$ and of the censoring distribution, respectively, for updating the Bayesian estimate $\hat{V}_{t_i}(t^*)$ at time t_i of interim analysis, as described

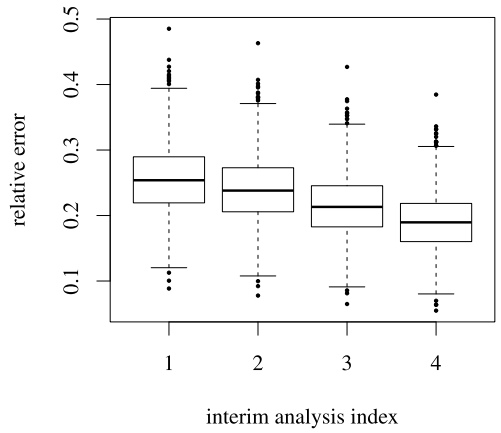

Figure 2. Box-plots of relative errors of $\hat{V}_{t_i}(t^*)$.

in Section 3.4. To see how $\hat{V}_{t_i}(t^*)$ improves with increasing i that accumulates more data, one can consider the relative error $|\hat{V}_{t_i}(t^*) - V(t^*)|/V(t^*)$. Figure 2 gives the box-plot of these relative errors over the 10,000 simulations for the ith interim analysis, $i = 1, 2, 3, 4$, in case 1 ($F = G$). It shows the decreasing trend of these relative errors with i. For the other cases, even though the relative errors become larger, with mean values ranging between 0.3 and 0.4, their decreasing trend with i are similar, showing that the estimates become more accurate with more data.

5. DISCUSSION

Unlike efficacy stopping, early stopping for futility during interim analyses does not inflate the type I error of the reference test. On the other hand, it may substantially decrease the power, or equivalently, substantially inflate the type II error if it is not carried out carefully. In particular, we have shown that the conditional power approach may have substantial power loss when it does not take into consideration the uncertainty of the estimated alternative. Whereas efficacy stopping is associated with rejection of $H_0 : \theta \le \theta_0$ at interim analysis, we can view futility stopping as rejection of the hypothesis $H_{\theta(M)} : \theta \ge \theta(M)$ associated with the implied alternative $\theta(M)$. This viewpoint enables us to use group sequential testing theory to develop a new approach to futility stopping. Simulation studies and asymptotic theory presented herein have shown its advantages over traditional conditional power and predictive power approaches, especially at the marginal alternatives.

Time sequential trials with survival outcomes have posed additional challenges to futility stopping. Lan and DeMets [15] have noted the difficulties due to the two time scales in these trials, namely *calendar time* and *information time*. The modified Haybittle-Peto approach we use in Section 3 circumvents the difficulties they cause for the commonly used error spending approach [12, Chapter 7]. As noted by Lin, Yao and Ying [18] for the logrank test, an additional challenge besides these two time scales is that the number of failure events observed during the course of the trial can substantially differ from that assumed at the design stage for determining the sample size and study duration of the trial, and it is useful to re-estimate this number, or more generally the maximum information $V(t^*)$ in the setting of Section 3.4, for futility decisions at interim analysis. Another useful contribution of the paper, therefore, is the new method proposed in Section 3.4 for updating the estimates of $V(t^*)$.

ACKNOWLEDGEMENTS

T. L. Lai's research was supported by the National Science Foundation grant DMS-085879 and the National Cancer Institute grant 1 P30 CA 124435-01. The research of Pei He and Olivia Liao was supported by the National Institutes of Health grant 4R37EB002784.

Received 25 October 2011

REFERENCES

[1] Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Process*. Springer, New York. MR1198884

[2] Bartroff, J., and Lai, T. L. (2008a). Efficient adaptive designs with mid-course sample size adjustment in clinical trials. *Stat. Med.* **27**, 1593–1611. MR2420330

[3] Bartroff, J., and Lai, T. L. (2008b). Generalized likelihood ratio statistics and uncertainty adjustments in efficient adaptive design of clinical trials. *Sequential Anal.* **27**, 254–276. MR2446902

[4] Chernoff, H., and Savage, R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Ann. Math. Statist.* **29**, 972–994. MR0100322

[5] Choi, S. C., Smith, P. J., and Becker, D. P. (1985). Early decision in clinical trials when treatment differences are small. *Contr. Clin. Trials.* **6**, 280–288.

[6] Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method. I. *Scand. J. Statist.* **16**, 97–128. MR1028971

[7] Gu, M., and Lai, T. L. (1998). Repeated significance testing with censored rank statistics in interim analysis of clinical trials. *Statistica Sinica* **8**, 411–428. MR1624347

[8] Halperin, M., Lan, K. K. G., Ware, J. H., Johnson, N. J., and DeMets, D. L. (1982). An aid to data monitoring in long-term clinical trials. *Contr. Clin. Trials.* **3**, 311–323.

[9] Herson, J. (1979). Predictive probability early termination plans for Phase II clinical trials. *Biometrics* **35**, 775–783.

[10] Hoeffding, W. (1960). Lower bounds for the expected sample size and the average risk of a sequential procedure. *Ann. Math. Statist.* **31**, 352–368. MR0120750

[11] Jennison, C., and Turbull, B. W. (1997). Group-sequential analysis incorporating covariate information. *J. Amer. Statist. Assoc.* **92**, 1330–1341. MR1615245

[12] Jennison, C., and Turbull, B. W. (2000). *Group Sequential Methods* **1**. Chapman & Hall, New York.

[13] Lai, T. L. (1975). Cherno-Savage statistics and sequential rank tests. *Ann. Statist.* **3**, 825–845. MR0388615

[14] Lai, T. L., and Shih, M. (2004). Power, sample size and adaptation considerations in the design of group sequential clinical trials. *Biometrika* **3**, 57–528. MR2090619

[15] Lan, K. K. G., and DeMets, D. L. (1989). Group sequential procedures: Calendar versus information time. *Statist. Med.* **8**, 1191–1198.

[16] Lan, K. K. G., Simon, R., and Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials theory. *Commun. Statist. C* **1**, 207–219. MR0685474
[17] Lan, K. K. G., and Wittes, J. (1988). The B-value: A tool for monitoring data. *Biometrics* **44**, 579–585.
[18] Lin, D. Y., Yao, Q., and Ying, Z. (1999). A general theory on stochastic curtailment for censored survival data. *Joul. Amer. Statist. Assoc.* **94**, 510–521. MR1702321
[19] Lorden, G. (1976). 2-SPRTs and the modied Kiefer-Weiss problem of minimizing an expected sample size. *Ann. Statist.* **4**, 281–291. MR0405750
[20] Pepe, M. S., and Anderson, G. L. (1992). Two-stage experimental designs: Early stopping with a negative result. *App. Statist.* **41**, 181–190.
[21] Spiegelhalter, D. J., Freedman, L. S., and Blackburn, P. R. (1986). Monitoring clinical trials: Conditional or predictive power? *Contr. Clin. Trials.* **7**, 8–17.
[22] Susarla, V., and Ryzin, J. V. (1976). Nonparametric bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.* **71**, 897–902. MR0436445
[23] Zhu, H., and Hu, F. (2010). Sequential monitoring of response-adaptive randomized clinical trials. *Ann. Statist.* **38**, 2218–2241. MR2676888

Pei He
390 Serra Mall Sequoia Hall
Stanford University, Stanford, CA 94305
USA
E-mail address: hepei@stanford.edu

Tze Leung Lai
390 Serra Mall Sequoia Hall
Stanford University, Stanford, CA 94305
USA
E-mail address: lait@stanford.edu

Olivia Y. Liao
390 Serra Mall Sequoia Hall
Stanford University, Stanford, CA 94305
USA
E-mail address: yuehwen@stanford.edu