Design of clinical trials with failure-time endpoints and interim analyses: An update after fifteen years

Pei He[a,*], Tze Leung Lai[b], and Zheng Su[c]

[a] *Genentech Inc., South San Francisco, CA 94080, USA*
[b] *Department of Statistics, Stanford University, Stanford, CA 94305, USA*
[c] *Deerfield Institute, New York, NY 10017, USA*

Abstract

Time to event is the clinically definitive endpoint in Phase III trials of new treatments of cancer, cardiovascular and many other diseases. Because these trials involve relatively long follow-up, their protocols usually incorporate periodic interim analyses of the data by a Data and Safety Monitoring Board/Committee. This paper gives a review of the major developments in the design of these trials in the 21st century, spurred by the need for better clinical trial designs to cope with the remarkable advances in cancer biology, genomics and imaging that can help predict patients' sensitivity or resistance to certain treatments. In addition to this overview and discussion of related issues and challenges, we also introduce a new approach to address some of these issues.

Keywords: Adaptive design, Calendar time, Early stopping, Multiple endpoints, Nonproportional hazards, Survival analysis

1. Introduction

Analysis of clinical studies with failure-time endpoints has been an important topic in biostatistics and has also led to a number of major methodological advances and important breakthroughs in statistical theory. A celebrated example is Cox's proportional hazards regression [1] that led to subsequent

*Corresponding author at: Genentech Inc., South San Francisco, CA 94080, USA, Tel: +1 6507992379
Email address: hepei1985@gmail.com (Pei He)

developments in partial likelihood, semiparametric efficiency, and statistical analysis of counting processes [2, 3, 4, 5, 6]. Although less renowned in comparison, the design of clinical trials with failure-time endpoints has also had important impact on clinical trial biostatistics and led to innovations in data monitoring and interim analysis of clinical trials. These innovations dated back to the seminal papers [7, 8] in 1982 on the Beta Blocker Heart Attack Trial (BHAT) and have continued until today, although at a much less spectacular pace than survival analysis. In this paper we give a review of the major developments in the 21st century, hence the "fifteen years" in the title. The "update" in the title refers to updating a previous review [9] that also provides a computer program to determine the power and sample size in the trial design; note that "design of clinical trials with failure-time endpoints and interim analyses" in our title is also the main part of the title of [9]. Since last year, we and our colleagues at Stanford University's Center for Innovative Study Design have been working to develop open-source software, which can be considered as an update of [9] for the implementation of some of the methods described in the next two sections. Section 2 reviews several new trends and innovative methods after the publication of [9] in the predecessor of this journal. In particular, it describes hybrid resampling for valid inference on primary and secondary endpoints of a survival trial, choice of stopping rules, adaptive designs including seamless phase II-III designs.

In his 2010 budget request, the Director of the National Cancer Institute earmarked "re-engineering" cancer clinical trials as a research initiative. The reason why re-engineering is needed is that although remarkable progress in biomedical sciences raised new hope for cancer treatment, the hope did not materialize because of the relatively small number of new anticancer agents that were demonstrated to be efficacious in phase III clinical trials, for which time to event (typically overall survival and occasionally progression-free survival) is a definitive endpoint. Besides choice of stopping boundaries, [9] also considers choice of test statistics. Being able to choose appropriate test statistics at terminal analysis can substantially increase the power of the commonly used logrank statistics in current designs that are mostly based on hazard ratios of treatment to control. In Section 3 we develop a new approach that allows adaptive choice of the test statistics at terminal analysis while still maintaining the prescribed Type I error. This circumvents one of the widely recognized difficulties with current survival trial designs that are dominated by hazard ratios and logrank statistics, which are inefficient for nonproportional hazards. The year 2010 also marked the appearance of the

much awaited FDA Draft Guidance for Industry on Adaptive Design. Two years later, the President's Council of Advisors on Science and Technology (PCAST) issued a report on "Propelling Inovations in Drug Discovery, Development, and Evaluation" and argued for using "innovative new approaches for trial design that can provide more information more quickly" as "it is increasingly possible to obtain clear answers with many fewer patients and with less time" by focusing studies on "specific subsets of patients most likely to benefit, identified based on validated biomarkers." Section 4 begins with a review of ongoing work in this direction for drug development and confirmatory testing, some of which is related to the approach introduced in Section 3. It then proceeds with further discussion and several concluding remarks.

2. Stopping rules, adaptive designs, and hybrid resampling

In Section 2.1 we review developments in the choice of stopping rules for time-sequential survival trials, in which "survival" refers to the failure-time endpoint in the title and "time-sequential" encapsulates the "interim analyses" that are carried out at prespecified calendar times. As pointed out in [10], survival trials have two time-scales - calendar time t and information time $V(t)$, which is the null variance of the test statistic at t. The information time $V(t)$ is the intrinsic time-scale for interim data but is typically unknown before time t unless restrictive assumptions are made *a priori*. In the past fifteen years, seamless Phase II-III designs and Bayesian adaptive designs are the most active area of research in innovative clinical trial designs. Section 2.3 gives a review of some of these developments in the context of time-sequential survival trials. Another important development in this period is hybrid resampling [11, 12, 13], which is reviewed in Section 2.2 and provides a basic tool in the methodological development in Section 3.

2.1. Early stopping for efficacy or futility at interim analysis

This basic problem in time-sequential survival trials is already addressed in [9], and we describe here subsequent developments. The censored rank

statistics considered in [9] and its precursors [14, 15] have the general form

$$S_n(t) = \sum_{i=1}^{n'} \delta_i'(t)\psi\left(H_{n,t}\left(X_i(t)\right)\right)\left\{1 - \frac{m_{n,t}'(X_i(t))}{m_{n,t}'(X_i(t)) + m_{n,t}''(X_i(t))}\right\}$$
$$- \sum_{j=1}^{n''} \delta_j''(t)\psi\left(H_{n,t}\left(Y_j(t)\right)\right)\frac{m_{n,t}''(Y_j(t))}{m_{n,t}'(Y_j(t)) + m_{n,t}''(Y_j(t))}, \quad (1)$$

where ψ is a nonrandom continuous function on $[0,1]$, $n = n' + n''$ is the total sample size, with n' patients assigned to treatment X and n'' assigned to treatment Y, $m_{n,t}'(s) = \sum_{i=1}^{n'} I\left(X_i(t) \geq s\right)$, $m_{n,t}''(s) = \sum_{j=1}^{n''} I\left(Y_j(t) \geq s\right)$, and $X_i(t)$, $Y_j(t)$, $\delta_i'(t)$, $\delta_j''(t)$ and $H_{n,t}(\cdot)$ are defined below. Let $T_i' \geq 0$ denote the entry time and $X_i > 0$ the survival time (or time to failure) after entry of the ith subject in treatment group X and let T_j'' and Y_j denote the entry time and survival time after entry of the jth subject in treatment group Y. The subjects are followed until they fail or withdraw from the study or until the study is terminated. Let ξ_i' (ξ_j'') denote the time to withdrawal, possibly infinite, of the ith (jth) subject in the treatment group X (Y). Thus the data at calendar time t consist of $(X_i(t), \delta_i'(t))$, $i = 1, \ldots, n'$, and $(Y_j(t), \delta_j''(t))$, $j = 1, \ldots, n''$, where $X_i(t) = \min\left(X_i, \xi_i', (t - T_i')^+\right)$, $\delta_i'(t) = I\left(X_i(t) = X_i\right)$, and $Y_j(t)$ and $\delta_j''(t)$ are defined similarly in terms of Y_j, ξ_j'' and T_j''. Let $H_{n,t}$ be the left-continuous version of the Kaplan-Meier estimator of the distribution function of the combined sample, defined by

$$1 - H_{n,t}(s) = \prod_{u<s}\left\{1 - \frac{\Delta N_{n,t}'(u) + \Delta N_{n,t}''(u)}{m_{n,t}'(u) + m_{n,t}''(u)}\right\}, \quad (2)$$

where $N_{n,t}'(s) = \sum_{i=1}^{n'} I\left(X_i \leq \xi_i' \wedge (t - T_i')^+ \wedge s\right)$, $N_{n,t}''(s) = \sum_{j=1}^{n''} I(Y_j \leq \xi_j'' \wedge (t - T_j'')^+ \wedge s)$, $\Delta N(s) = N(s) - N(s-)$ and we use the convention $0/0 = 0$. For the time-sequential censored rank statistics (1), Gu and Lai [14] showed that $\{S_n(t)/\sqrt{n},\ t \geq 0\}$ converges weakly to a Gaussian process with independent increments and variance function $V(t)$ under the null hypothesis $H_0 : F = G$ and contiguous alternatives. Two commonly used estimates of S_n are

$$V_n^1(t) = \int_0^t \frac{\psi^2(H_{n,t})(s)m_{n,t}'(s)m_{n,t}''(s)}{(m_{n,t}'(s) + m_{n,t}''(s))^2}\, d\left(N_{n,t}'(s) + N_{n,t}''(s)\right), \quad (3)$$

or

$$V_n^2(t) = \int_0^t \frac{\psi^2(H_{n,t}(s))}{(m'_{n,t}(s) + m''_{n,t}(s))^2} \left\{ \left(m''_{n,t}(s)\right)^2 \ dN'_{n,t}(s) + \left(m'_{n,t}(s)\right)^2 \ dN''_{n,t}(s) \right\}. \tag{4}$$

As a compromise between these two choices, Gu and Lai [14] also considered $V_n^3(t) = \{V_n^1(t) + V_n^2(t)\}/2$. For any choice $V_n(t)$ of the three estimates, $n^{-1}V_n(t)$ converges in probability to $V(t)$ under H_0 and under contiguous alternatives. When the patients are randomized to X or Y with probability $1/2$, $\gamma = 1/2$, $m'_{n,t} \sim m''_{n,t}$ under $F = G$. Therefore for the logrank statistic for which $\psi \equiv 1$, $V_n^1(t)$ and $V_n^1(t)$ are asymptotically equivalent to $V_n(t) =$ (total number of deaths up to time t)/4, which is the standard formula for the null variance estimate of the logrank statistic in randomized clinical trials. The program in [9] allows the user to choose ψ from Self's [16] beta family $\psi(u) = u^\rho(1-u)^\tau$ by specifying the values of $\rho \geq 0$ and $\tau \geq 0$. The case $\tau = 0$ yields the G^ρ statistics proposed by Harrington and Fleming [17], with $\rho = 0$ corresponding to the logrank statistic and $\rho = 1$ corresponding to the Peto-Prentice generalization of of Wilcoxon's statistic [18, 19].

With this choice of the test statistics $S_n(t)$, [9] allows the user to choose from three classes of stopping boundaries for either two-sided or one-sided tests using the normalized statistics $W_i = S_n(t_i)/\sqrt{V_n(t_i)}$, where $t_1, \ldots, t_k$ are the calendar times of interim analysis. The first class is referred to as Slud and Wei's boundaries [20]. It requires the user to specify positive numbers $\alpha_1, \ldots, \alpha_k$ such that $\sum_{j=1}^k \alpha_j = \alpha$. The stopping boundaries b_j can be determined recursively by

$$P_{F=G}\left\{ |W_1| \leq b_1\sqrt{V_1}, \ldots, \ |W_{j-1}| \leq b_{j-1}\sqrt{V_{j-1}}, \ |W_j| > b_j\sqrt{V_j} \right\} = \alpha_j, \tag{5}$$

in which the probability on the left-hand side can be computed by recursive numerical integration [21, Section 4.3.1] in view of the aforementioned asymptotic normally (with independent increments) of $(S_n(t_1), \ldots, S_n(t_k))$. However, there are no guidelines nor systematic ways to choose the user-specified α_j. The second class of stopping boundaries is called Lan-DeMets, named after the authors of [22] that introduced the "error spending" approach to specifying stopping boundaries. To apply the error spending approach to time-to-event responses, one needs an a priori estimate of the null variance of $S_n(t_k)$. Let v_1 be such an estimate. Although the null variance of $S_n(t)$ is

expected to be nondecreasing in t under the asymptotic independent increments property, its estimate $V_n(t)$ may not be monotone, and a simple fix is to redefine $V_n(t_j)$ to be $V_n(t_{j-1})$ if $V_n(t_j) < V_n(t_{j-1})$. Let $\pi : [0, v_1] \to [0, 1]$ be a nondecreasing function with $\pi(0) = 0$ and $\pi(v_1) = \alpha$, which can be taken as the error spending function of a stopping rule τ (taking values in $[0, v_1]$) of a Wiener process. Letting $\alpha_j = \pi(v_1 \wedge V_n(t_j)) - \pi(V_n(t_{j-1}))$ for $j < k$ and $\alpha_k = \alpha - \pi(V_n(t_{k-1}))$, the boundary values $b_1, \ldots, b_K$ are defined recursively by (5), in which $\alpha_j = 0$ corresponds to $b_j = \infty$. This test has type I error probability approximately equal to α, irrespective of the choice of π and the *a priori* estimate v_1. Its power, however, depends on π and v_1. The requirement that the trial be stopped once $V_n(t)$ exceeds v_1 is a major weakness of the preceding stopping rule. Since one usually does not have sufficient prior information about the underlying survival distributions and the actual accrual rate or the withdrawal pattern, v_1 may substantially over- or under-estimate the expected value of $V_n(t_k)$. Scharfstein, Tsiatis and Robins [23, 24] have proposed re-estimation procedures during interim analyses to address this difficulty, but re-estimation raises concerns about possible inflation of the type I error probability.

The third class of stopping boundaries is called "modified Haybittle-Peto" (modHP) and depends on a user-specified value of b for $b_1 = \cdots = b_{k-1} = b$ in (5) so that $b_k = c$ can be determined by

$$P\left\{|W\left(V_n(t_j)\right)| \geq bV_n^{1/2}(t_j) \text{ for some } j < k \right. \\ \left. \text{or } |W\left(V_n(t_k)\right)| \geq cV_n^{1/2}(t_k) \,\middle|\, V_n(t_1), \ldots, V_n(t_k)\right\} = \alpha, \quad (6)$$

which is in fact a modification of an earlier proposal by Haybittle [25] who considered asymptotically normal test statistics $S_n(t_i)$ that behave like a normal random walk. Haybittle used some relatively large value of b, such as 3, and conventional critical values of c for the final analysis at t_k in case the trial does not stop at the earlier analyses. Peto et al. [26] subsequently advocated to use Haybittle's design for randomized clinical trials that require "prolonged observation for each patient." Gu and Lai [15] proposed to determine c by (6) to guarantee the prescribed level α for the type I error, and noted that a major advantage of modHP over the error spending approach is that it does not require an *a priori* estimate of $V_n(t_k)$ at interim analyses prior to t_k. Six years later, Lai and Shih [27] developed a theory of group sequential tests of the one-sided hypothesis $H_0 : \theta \leq \theta_0$, for the parameter

θ of an exponential family of densities $e^{\theta z-\phi(z)}$ (which includes the case of normal densities with known variance as a special case), with significance level α and a maximum number M of observations. This theory shows that the modHP test has nearly optimal power and expected sample size under the constraints α and M, which [27] also confirms in simulation studies that show in particular its superiority over the error-spending approach, even in this case of v_1 being proportional to M.

2.2. Hybrid resampling and inference in time-sequential designs

A general framework for statistical inference, in particular the problem of constructing confidence intervals in sequential experiments, was introduced fifteen years ago by Chuang and Lai [11]. Let $\mathbf{X}$ be a vector of observations from some family of distributions $\{F : F \in \mathscr{F}\}$. For nonparametric problems, $\mathscr{F}$ is the family of distributions satisfying certain prespecified regularity conditions. For parametric models with parameter $\eta \in \Gamma$, we can denote $\mathscr{F}$ by $\{F_\eta : \eta \in \Gamma\}$. The problem of interest is to construct a confidence interval for the real-valued parameter $\theta = \theta(F)$, and we next review these methods for the construction. Let Θ denote the set of all possible values of θ.

Exact method: If $\mathscr{F} = \{F_\theta : \theta \in \Theta\}$ is indexed by a real-valued parameter θ, an exact equal-tailed confidence region can always be found by using the well-known duality between hypothesis tests and confidence regions. Suppose one would like to test the null hypothesis that θ is equal to θ_0. Let $R(\mathbf{X}, \theta_0)$ be some real-valued test statistic. Let $u_\alpha(\theta_0)$ be the α-quantile of the distribution of $R(\mathbf{X}, \theta_0)$ under the distribution F_{θ_0}. The null hypothesis is accepted if $u_\alpha(\theta_0) < R(\mathbf{X}, \theta_0) < u_{1-\alpha}(\theta_0)$. An exact equal-tailed confidence region with coverage probability $1 - 2\alpha$ consists of all θ_0 not rejected by the test and is therefore given by $\{\theta : u_\alpha(\theta) < R(\mathbf{X}, \theta) < u_{1-\alpha}(\theta)\}$. This method applies only when there are no nuisance parameters.

Bootstrap method: The bootstrap method replaces the quantiles $u_\alpha(\theta)$ and $u_{1-\alpha}(\theta)$ by the following approximate quantiles u^*_α and $u^*_{1-\alpha}$. Let $\hat{F}$ be an estimate of $F \in \mathscr{F}$ based on $\mathbf{X}$. The quantile u^*_α is defined to be α-quantile of the distribution of $R(\mathbf{X}^*, \hat{\theta})$ with $\mathbf{X}^*$ generated from $\hat{F}$ and $\hat{\theta} = \theta(\hat{F})$. This yields the confidence region $\left\{\theta : u^*_\alpha < R(\mathbf{X}, \theta) < u^*_{1-\alpha}\right\}$ for θ with approximate coverage probability $1 - 2\alpha$. In particular, when $\hat{F}$ is the empirical distribution of i.i.d. $X_1, \ldots, X_n$ and $R(\mathbf{X}, \theta) = (\hat{\theta} - \theta)/\hat{\sigma}$ for some estimate $\hat{\sigma}$ of the standard error of $\hat{\theta}$, the bootstrap confidence interval is called the bootstrap-t interval.

Hybrid resampling method: The hybrid confidence region is based on reducing the family of distributions $\mathscr{F}$ to another family of distributions $\{\hat{F}_\theta : \theta \in \Theta\}$, which is used as the "resampling family" and in which θ is the unknown parameter of interest. Let $\hat{u}_\alpha(\theta)$ be the α-quantile of the sampling distribution of $R(\mathbf{X}, \theta)$ under the assumption that X has distribution $\hat{F}_\theta$. The hybrid confidence region results from applying the exact method to $\{\hat{F}_\theta : \theta \in \Theta\}$ and is given by

$$\{\theta : \hat{u}_\alpha(\theta) < R(\mathbf{X}, \theta) < \hat{u}_{1-\alpha}(\theta)\}. \tag{7}$$

The construction of (7) typically involves simulations to compute the quantiles as in the bootstrap method, and is called the *hybrid resampling* method because it "hybridizes" the exact method (that uses test inversion) with the bootstrap method (that uses the observed data to determine the resampling distribution). Note that hybrid resampling is a generalization of the bootstrap method, which uses the singleton $\{\hat{F}\}$ as the resampling family $\{\hat{F}_\theta\}$. In practice, it is often desirable to express a confidence set for θ as an interval. Although (7) may not be an interval, it often suffices to give only the upper and lower limits of the confidence set. An algorithm, based on method of successive secant approximations, is given in [11] to find the upper or lower limit of (7).

Since an exact or hybrid resampling method for constructing confidence regions is based on inverting a test, it is implicitly or explicitly linked to an ordering of the sample space of the test statistic used. The ordering defines the p-value of a test as the probability (under the null hypothesis) of more extreme values (under the ordering) of the test statistic than that observed in the sample. Equivalently, the test rejects the null hypothesis, one for each given θ, if the test statistic exceeds or falls below a specified quantile of its null distribution. Let $X_1, X_2, \ldots$ be i.i.d. random variables, $S_n = X_1 + \cdots + X_n$, and T be a stopping time. Under a total ordering $\leq$ of the sample space of (T, S_T), Lai and Li [12] call (t, s) a qth quantile if $P\{(T, S_T) \leq (t, s)\} = q$, assuming that the X_i have a strictly increasing continuous distribution function. This is a natural generalization of the qth quantile of a univariate random variable. For randomly stopped sums of independent normal random variables with unknown mean θ, the bivariate vector (T, S_T) is sufficient for θ. For the general setting where a stochastic process $\mathbf{X}_u$, in which u denotes either discrete or continuous time, is observed up to a stopping time T, [12] defines $\mathbf{x} = (x_u, u \leq t)$ to be a qth quantile if

$P\{\mathbf{X} \le \mathbf{x}\} \ge q$ and $P\{\mathbf{X} \ge \mathbf{x}\} \ge 1-q$, under a total ordering $\le$ for the sample space of $\mathbf{X} = (\mathbf{X}_u, u \le T)$.

For applications to confidence intervals of a real parameter θ, the choice of the total ordering should be targeted towards the objective of interval estimation. Let U_r $(r \le T)$ be real-valued statistics based on the observed process $\mathbf{X}_s (s \le T)$. For example, let U_r be an estimate of θ based on $\{\mathbf{X}_s, s \le r\}$. A total ordering on the sample space of X can be defined by

$$\mathbf{X} \ge \mathbf{x} \quad \text{if and only if} \quad U_{T \wedge t} \ge u_{T \wedge t}, \tag{8}$$

where $T \wedge t = \min(T, t)$ and $(u_r, r \le t)$ is defined from $\mathbf{x} = (\mathbf{x}_r, r \le t)$ in the same way as $(U_r, r \le T)$ is defined from X. In particular, consider the case of independent normal X_n and let U_n be the sample mean $\bar{X}_n$ of $X_1, \ldots, X_n$. In this case, (8) yields the ordering

$$(T, S_T) \ge (t, s_t) \quad \text{if and only if} \quad \bar{X}_{T \wedge t} \ge s_{T \wedge t}/(T \wedge t),$$

which is equivalent to the ordering scheme introduced by Siegmund [28] to construct exact confidence intervals for the mean of a normal distribution with known variance in a sequential design. Lai and Li [12] show how the ordering scheme (8) can be applied in conjunction with hybrid resampling to construct confidence intervals of the hazard ratio following time-sequnetial tests in the proportional hazards model. Lai, Shih and Su [13] apply (8) to construct hybrid resampling confidence intervals for secondary endpoints in group sequential or time-sequential trials for which the stopping rule is based on a primary endpoint, an example of which is the commonly used hazard ratio in survival trials.

2.3. Bayesian approach and adaptive seamless designs of time-sequential survival trials

As pointed out in [8], BHAT was actually designed as a fixed-duration (instead of time-sequential) trial. It was stopped early by the Data and Safety Monitoring Board eight months before the prescheduled of the trial by arguments involving stochastic curtailment of the fixed-duration trial. The basic idea is to stop a nonsequential level-α test of $\mathrm{H}_0 : \theta = \theta_0$ if the conditional power $p_t(\theta') = P_{\theta'}(\text{Reject } \mathrm{H}_0 | D_t)$ at a given alternative θ' given the data D_t up to time t falls below some threshold $1 - \rho_1$, resulting in the acceptance of H_0, or if the conditional type I error $P_{\theta_0}(\text{Reject } \mathrm{H}_0 | D_t)$ exceeds ρ_0, leading to rejection of H_0. It is shown in [29] that for i.i.d. normally distributed

observations with unknown mean θ and known variance, the curtailed test has type I error $\leq \alpha/\rho_0$ and type II error α'/ρ_1, where α' is the type II error (at θ') of the original nonsequential test. Lin, Yao and Ying [30] discuss several subtle issues in the definition and implementation of conditional power for censored survival data in time-sequential trials. One issue is that the conditional distribution of $S_n(t^*)$ given $S_n(t)$ may not be the same as that given $D(t)$, where $S_n(t)$ is the censored rank statistic (1) and t^* denotes the prescheduled termination time of the trial. Another issue is that the actual accrual, failure and censoring patterns may differ substantially from those anticipated at the design stage, making it necessary to re-evaluate certain quantities in the weak convergence theory of $\{S_n(t), t \leq t^*\}$, under the null hypothesis and contiguous alternatives, which they use for the implementation of stochastic curtailment.

Conditional power and stochastic curtailment basically involve prediction of $S_n(t^*)$, under the null hypothesis and a specified alternative (which is used for sample size determination at the design stage), given all the data up to the time t of interim analysis. Spigelhalter, Freedman and Blackburn [31] advocated to use a Bayesian approach for such prediction instead. Assuming a prior distribution on θ, Bayesian prediction is based on the posterior distribution $\pi(\theta|D_t)$ and the predictive power is defined by

$$P_t = P(\text{Reject } H_0|D_t) = \int p_t(\theta) d\pi(\theta|D_t). \tag{9}$$

For censored survival data from time-sequential trials, the posterior distribution $\pi(\theta|D_t)$ is complicated but can be evaluated by Markov Chain Monte Carlo (MCMC) methods when parametric or semiparametric models are assumed on the survival distributions, with θ representing the parameter vector; see [32, 33], [34, Section 4], [35, Sections 5.6–5.8]. As pointed out in Section 2.5 of [35], the Bayesian approach to stochastic curtailment uses predictive power instead of conditional power. This approach to early stopping of a clinical trial for futility or efficacy does not have type I error probability guarantees. Acknowledging that type I error probability guarantees are important to gain regulatory approval of a new treatment, "frequentist twists" are used to satisfy the type I error constraint at some chosen parameter configuration in the null hypothesis by Monte Carlo simulations at the configuration and thereby adjusting the rejection threshold of the Bayesian test [35, 36]. However, as pointed out in [37], there is no guarantee that the type I error is maintained at other parameter configurations for a composite null

hypothesis, as in semiparametric models for survival outcomes. In contrast, the modified Haybittle-Peto stopping boundary applied to commonly used censored rank statistics in comparing the survival distributions of two treatments has frequentist validity besides efficiency, flexibility and ease of use in time-sequential survival trials, showing its advantages over the preceding Bayesian approach to early stopping.

The Bayesian approach, however, has been applied to develop innovative clinical trial designs for much more complex settings involving survival outcomes than early stopping before the prescheduled termination date as in BHAT, which was the focus of [9]. One such setting is adaptive design of Phase II-III oncology trials. The majority of Phase II studies in oncology leading to Phase III clinical trials are are single-arm studies with a binary tumor response endpoint and the most commonly used phase II designs are Simon's [38] single-arm two-stage designs for testing $H_0 : p \leq p_0$ versus $H_1 : p \geq p_1$ where p is tumor response rate. Whether the new treatment is declared promising in a single-arm Phase II trial, however, depends strongly on the prespecified p_0 and p_1. As noted by Vickers et al. [39], uncertainty in the choice of p_0 and p_1 can increase the likelihood that (a) a treatment with no viable positive treatment effect proceeds to Phase III, for example, if p_0 is chosen artificially small to inflate the appearance of a positive treatment effect when one exists, or (b) a treatment with positive treatment effect is prematurely abandoned at Phase II, for example, if p_1 is chosen optimistically large. To circumvent the problem of choosing p_0, [39] and [40] have advocated randomized Phase II designs. In particular, it is argued that randomized Phase II trials are needed before proceeding to Phase III trials when (a) there is not a good historical control rate, due to either incomparable controls (causing bias), few control patients (resulting in large variance of the control rate estimate), or outcome that is not "antitumor activity", or when (b) the goal of Phase II is to select one from several candidate treatments or several doses for use in Phase III. However, few Phase II cancer studies are randomized with internal controls. The major barriers to randomization include that randomized designs typically require a much larger sample size than single-arm designs and that there are multiple research protocols competing for a limited patient population. Being able to include the Phase II study as an internal pilot for the confirmatory Phase III trial may be the only feasible way for a randomized Phase II cancer trial of such sample size and scope to be conducted.

Although tumor response is an important treatment outcome, the clin-

ically definitive endpoint in Phase III cancer trials is usually time to event (death or recurrence). The go/no-go decision to Phase III is typically based on tumor response because the clinical time-to-failure endpoints in Phase III are often of long latency. Seamless Phase II-III trials with bivariate endpoints consisting of tumor response and time to event are an important accomplishment of the Bayesian approach, introduced by Inoue, Berry and Thall [41] and Huang et al. [42] to relate survival to response. Let z_i denote the treatment indicator (0=control, 1=experimental), τ_i denote survival time, and y_i denote the binary response for patient i. The Bayesian approach assumes that the responses y_i are independent Bernoulli variables and the survival time τ_i given y_i follows an exponential distribution, denoted $\text{Exp}(\lambda)$ in which $1/\lambda$ is the mean: $y_i \mid z_i = z \overset{\text{i.i.d.}}{\sim} \text{Bern}(\pi_z)$, $\tau_i \mid \{y_i = y,\ z_i = z\} \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda_{z,y})$. It then follows that the conditional distribution of τ_i given z_i is a mixture of exponentials:

$$\tau_i \mid z_i = z \overset{\text{i.i.d.}}{\sim} \pi_z \text{Exp}(\lambda_{z,1}) + (1 - \pi_z)\text{Exp}(\lambda_{z,0}). \tag{10}$$

Instead of the parametric assumption of $\text{Exp}(\lambda_{z,y})$ for the conditional distribution of τ_i, semiparametric methods such as Cox regression, however, are often preferred for reproducibility considerations and because of the relatively large sample sizes in phase III studies. This led Lai, Lavori and Shih [43] to develop an alternative seamless phase II-III design that uses a semiparametric model to relate survival to response and is directly targeted toward frequentist testing with generalized likelihood ratio (GLR) or partial likelihood statistics. Their basic idea is to replace the stringent parametric model involving exponential distributions by a semiparametric counterpart that generalizes the Inoue-Thall-Berry model. Let y denote the response and z denote the treatment indicator, taking the value 0 or 1. Consider the proportional hazards model $\lambda(t \,|\, y, z) = \lambda_0(t) \exp(\alpha y + \beta z + \gamma yz)$. The Inoue-Thall-Berry exponential model is a special case with $\lambda_0(\cdot)$ being the constant hazard rate of an exponential distribution. Let $\pi_0 = \text{pr}(y = 1 \,|\, \text{control})$ and $\pi_1 = \text{pr}(y = 1 \,|\, \text{treatment})$. Let $a = e^\alpha$, $b = e^\beta$ and $c = e^\gamma$, and let S be the survival function and f be the density function associated with the hazard function λ_0 so that $\lambda_0 = f/S$. In this augmented proportional hazards model, the survival distribution of τ is

$$\text{P}(\tau > t) = \begin{cases} (1 - \pi_0)S(t) + \pi_0(S(t))^a & \text{for the control group } (z = 0), \\ (1 - \pi_1)(S(t))^b + \pi_1(S(t))^{abc} & \text{for the treatment group } (z = 1). \end{cases} \tag{11}$$

The hazard ratio of the treatment to control survival varies with t because of the mixture form in (11). Let $\boldsymbol{\pi} = (\pi_0, \pi_1)$, $\boldsymbol{\xi} = (a, b, c)$. Lai, Lavori and Shih [43] formulate the null hypothesis as $H_0 : \pi_0 \geq \pi_1,$ or $\pi_0 < \pi_1$ and $d(\boldsymbol{\pi}, \boldsymbol{\xi}) \leq 0$, and use time-sequential GLR and partial likelihood ratio statistics to test H_0, where $d(\boldsymbol{\pi}, \boldsymbol{\xi})$ is the limiting hazard ratio (which does not depend on t) as $a \to 1$ and $c \to 1$.

3. A new approach to the design and analysis of time-sequential survival trials

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. 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. 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 may not actually hold. In this case, as the sample size n approaches ∞, the limiting distribution of $S_n(t)/\sqrt{n}$ 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 $S_n(t)/\sqrt{n}$ may not be linear in $V(t)$, and may level off or even decrease with increasing $V(t)$, as shown in [15].

In the past decade, the logrank statistic and the closely related hazard ratio together with the proportional hazards model have become the most widely used test statistic and endpoint for survival trials. Their popularity in the medical literature is partly due to the conceptual simplicity of the hazard ratio as a summary measure to compare two survival distributions, whose Kaplan-Meier estimates are typically also plotted in the report of a survival trial. Although the logrank statistic $S_n(t)$ is indeed asymptotically efficient under the proportional hazards model, it can have a flattening or

eventually decreasing drift $\sqrt{n}\mu(t)$ under local alternatives, as pointed out above. In this case, some other functionals of the survival distributions are more appropriate than the hazard ratio which is no longer constant without the proportional hazards assumption. These functionals are given in Section 3.1 in which we describe how they can be used in conjunction with the logrank statistic for the analysis at the prescheduled end $t^*(= t_k)$ of a time-sequential survival trial whose early stopping rule involves the logrank statistic and the modified Haybittle-Peto boundary. Thus, at the planning stage and at interim analyses, we follow the popular practice of carrying out repeated logrank tests. Even when interim data show clear departures from the proportional hazards model, early efficacy stopping of the time-sequential logrank test may have both increase in power and reduction in sample size if the mean $\mu(t)$ of $S_n(t)/\sqrt{n}$ decreases with increasing $V(t)$, as shown in [15]. What can really hurt by using the inefficient logrank statistic in such cases is when early stopping has not occurred during interim analyses and one can end up with substantial loss of power with the logrank test at the prescheduled end t^* of the trial. However, our new approach salvages the power loss by bringing in other functionals of the survival descriptions. This approach makes use of certain ideas in Section 2.2 for its implementation. Section 3.2 presents simulation results on the performance of this new design.

Whereas Section 3.1 focuses on early stopping for efficacy and terminal analysis at t^*, we consider early stopping for futility in Section 3.3. He, Lai and Liao [44] have recently developed a theory for futility stopping using the idea of an "implied alternative" introduced by Lai and Shih [27] in connection with the efficiency theory of modified Haybittle-Peto tests. In time-sequential survival trials, this implied alternative depends on $V(t^*)$, which [44] uses a Bayesian approach to estimate during the course of the trial. Combining Sections 3.1 and 3.3 yields a new time-sequential design that updates previous works in Section 2.1.

3.1. Cumulative hazard differences to supplement the logrank statistic at t^*

As noted above, when the trial results are published to show the survival benefits of the new treatment, the Kaplan-Meier curves of the treatment and control groups are usually included in the report. We consider here the possibility of combining some key features of the survival curves with the logrank statistic at the prescheduled end of the trial to enhance the power of the test in the case of non-proportional hazards. Instead of Kaplan-Meier curves, it may be more convenient to consider the Studentized

cumulative hazard differences at selected survival times $s_1, \ldots, s_L$ (e.g., 1-year, 2-year, 5-year survival). Using the same notation as in Section 2.1, let $\hat{\Lambda}_X(s) = \Sigma_{u \le s} \left(\Delta N'_{n,t^*}(u) / m'_{n,t^*}(u) \right)$, $\hat{\Lambda}_Y(s) = \Sigma_{u \le s} \left(\Delta N''_{n,t^*}(u) / m''_{n,t^*}(u) \right)$ be the Nelson-Aalen estimators of the cumulative hazard functions of the two groups at the prescheduled termination time t^* of the trial. Let $V_X(s) = \Sigma_{u \le s} \Delta N'_{n,t^*}(u) / (m'_{n,t^*}(u))^2$ be the estimate of $\mathrm{Var}(\hat{\Lambda}_X(s))$ and define $V_Y(s)$ similarly. Define the Studentized cumulative hazard difference at s_l ($l = 1, \ldots, L$) by

$$\Delta_l = \left(\hat{\Lambda}_X(s_l) - \hat{\Lambda}_Y(s_l) \right) / \left(V_X(s_l) + V_Y(s_l) \right)^{1/2}. \tag{12}$$

Instead of the estimated cumulative hazards, one can use the Kaplan-Meier estimates $\hat{S}_X$ and $\hat{S}_Y$ of the survival functions at s_l that are usually graphed in reporting the trial results, replacing $\hat{\Lambda}_X(s_l) - \hat{\Lambda}_Y(s_l)$ by $\hat{S}_X(s_l) - \hat{S}_Y(s_l)$ in (12) so that $V_X(s_l)$ is now given by the Greenwood formula

$$\hat{S}_X^2(s_l) \Sigma_{u \le s} \Delta N'_{n,t^*}(u) / \{ m'_{n,t^*}(u) [m'_{n,t^*}(u) - \Delta N'_{n,t^*}(u)] \},$$

and $V_Y(s_l)$ is given by its corresponding Greenwood formula. We call this modification the "survival variant" of (12).

The test statistics $\Delta_1, \ldots, \Delta_L$ defined by (12) or its survival variant are used to supplement the Studentized logrank statistic $S_n(t^*)/\hat{\sigma}_n(t^*)$, where $\hat{\sigma}_n^2(t) =$ (total number of deaths up to time t)/4. Thus, the time-sequential test statistics W_i now take the form

$$W_i = \begin{cases} S_n(t_i)/\hat{\sigma}_n(t_i) & \text{for } i \le i < k, \\ \max(S_n(t^*)/\hat{\sigma}_n(t^*), \Delta_1, \ldots, \Delta_L) & \text{for } i = k. \end{cases} \tag{13}$$

We apply the modified Haybittle-Peto stopping rule in Section 2.1 to the test statistics (13), in which $S_n(t)$ is the logrank statistic at calendar time t.

3.2. Implementation and a simulation study

The determination of b in the modified Haybittle-Peto test proceeds as in [9, 15] for the symmetric boundaries and as in [27, 44] for asymmetric boundaries b and $\tilde{b}$. The determination of c becomes much more difficult after introducing the additional test statistics $\Delta_1, \ldots, \Delta_L$ in (13) because $(\Delta_1, \ldots, \Delta_L, S_n(t_i)/\hat{\sigma}_n(t_i), 1 \le i \le k)$ does not have an independent increments correlation structure. Instead of multivariate integration after applying the joint limiting normal distribution of the random vector under

$F = G$ to evaluate the probability in (6), we compute the probability by Monte Carlo simulations, using the procedure in [12, p.644]. Moreover, the threshold c in the modified Haybittle-Peto test does not need to be computed explicitly because checking whether the observed values W_k^{obs} of W_k exceeds c if the trial has not stopped prior to t_k is equivalent to checking whether $\hat{P}\{(t_{i^*}, W_{i^*}^*) > (t_k, W_k^{\text{obs}})\} \leq \alpha$, where the ordering $>$ is the Lai-Li ordering (8), i^* represents the time index of the interim analysis at which the modified Haybittle-Peto test based on the random variables W_i^* (generated from $\hat{P}$) stops, and $\hat{P}$ is the probability measure corresponding to the estimated common survival distribution of the two groups.

We apply this procedure to implement the proposed design in the following comparative study of its performance. The study simulates a clinical trial that enrolls 450 patients uniformly over a 6-year period. Interim analyses are performed at $t = 2, 4, 6$ years, and the prescheduled end of the trial is $t^* = 8$ years. There are four scenarios in the simulation study, in which Y is exponential with mean 3 and represents the survival time of a patient drawn at random from the control group.

Scenario A: X has the same distribution as Y, representing a particular model in the composite null hypothesis.

Scenario B (proportional hazards): Hazard ratio of X to Y is 2/3.

Scenario C: The hazard ratio of X to Y varies with the survival time s and is 0.4 for $s \leq 0.8$ and increases to 1 for $s > 0.8$.

Scenario D: The hazard ratio of X to Y is 0.4 for $1 \leq s \leq 3$ and is 1 elsewhere.

The proposed design, labeled Design 1, or Design 2 for its survival variant, is compared with three other designs described below. Design 1 (or Design 2) chooses $L = 2$, $s_1 = 1$ and $s_2 = 3$, and therefore combines the Studentized cumulative hazard differences (or Studentized survival differences) at 1 and 3 years with the Studentized logrank statistic at terminal analysis. Designs 3, 4 and 5 involve only the time-sequential logrank statistics, and use the modified Haybittle-Peto boundary and the Pocock and O'Brien-Fleming stopping rules (as in Example 2 of [15]) respectively. Table 1 gives the type I error and power $P(\text{Reject } \text{H}_0)$ of the five designs in each scenario. Each result is based on 2000 simulations. The table shows that all five designs maintain the type I error of 0.05 (Scenario A), and that Designs 1 and 2, have power comparable to the other three designs in the proportional hazards model (Scenario B), but are substantially more powerful in the non-proportional settings of Scenarios C and D. Since Designs 1 and 2 use the same stopping rule at the interim

analyses as Design 3 whose expected study duration performance has been more extensively studied in [15], we do not include in Table 1 results on expected durations. Figure 1 plots the survival distributions of the treatment and control groups for Scenarios B, C and D.

3.3. Bayesian prediction of future $V_n(t)$

The Bayesian prediction approach in [44] to estimating at time t_i the null variance $V_n(t)$ of the score statistic $S_n(t)$ for $t > t_i$ uses 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 κ times the assumed parametric model, which is typically a proportional hazards model, used for power calculation at the design stage, where $\kappa = \alpha(\mathbb{R}_+)$ 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 $Z_j^{(i)}, j = 1, \ldots, n_i$, denote the combined sample of $X_l(t_i), Y_h(t_i)$, using the same notation as that in Section 2.1. Let $\delta_j^{(i)}$ be the censoring indicator $\delta_l'(t_i)$ or $\delta_h''(t_i)$ associated with $Z_j^{(i)}$, and $m_i(s)$ be the corresponding $m_{n,t}'(s)$ or $m_{n,t_i}''(s)$. By re-arranging the observations, assume without loss of generality that $Z_1^{(i)}, \ldots, Z_k^{(i)}$ are the uncensored observations, and let $Z_{[k+1]}^{(i)} < \cdots < Z_{[m]}^{(i)}$ denote the distinct ordered censored observations. Let $m_i^+(s) = \sum_{j=1}^{n_i} I_{\{Z_j^{(i)} > s\}}$, $\lambda_i(s) = \sum_{j=1}^{n_i} I_{\{Z_j^{(i)} = s, \delta_j = 0\}}$, $Z_{[k]}^{(i)} = 0$, $Z_{[m+1]}^{(i)} = \infty$. As shown in [45], for $Z_{[l]}^{(i)} \le u < Z_{[l+1]}^{(i)}$, the Bayes estimate of $H(u)$ at the ith interim analysis is given by

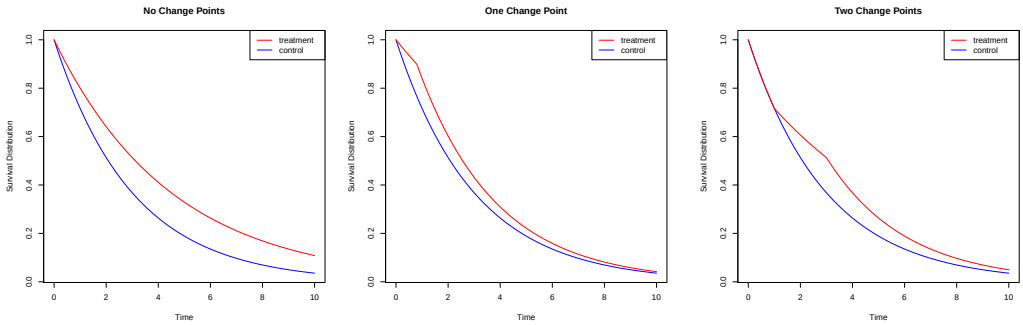
$$\hat{H}_i(u) = \frac{\alpha(u, \infty) + m_i^+(u)}{\alpha(\mathbb{R}_+) + n_i} \times \prod_{j=k+1}^{l} \left\{ \frac{\alpha[Z_{[j]}^{(i)}, \infty) + N_i(Z_{[j]}^{(i)})}{\alpha[Z_{[j]}^{(i)}, \infty) + m_i(Z_{[j]}^{(i)}) - \lambda_i(Z_{[j]}^{(i)})} \right\}.$$

Similarly, for updating the estimate $\hat{C}$ of the censoring distribution, [44] interchanges the roles of τ_j and ξ_j above and replaces α by α_c that is associated with the specification of the censoring distribution at the design

Table 1: P(Reject H_0) for five designs in Scenarios A, B, C and D.

Scenario	Design				
	1	2	3	4	5
A	0.047	0.048	0.049	0.050	0.049
B	0.905	0.890	0.932	0.902	0.939
C	0.853	0.858	0.658	0.672	0.651
D	0.671	0.707	0.522	0.395	0.543

stage. At the ith interim analysis, the accrual rates for the period prior to t_i have been observed and those for the future 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 $(Z_j^{(i)}, \delta_j^{(i)})$ observed up to time t_i.



(a) Scenario B **(b)** Scenario C **(c)** Scenario D

Figure 1: Survival distributions of the treatment and control groups

4. Discussion

4.1. Flexibility and efficiency of the new approach to early stopping

Combining the commonly used logrank statistics with other statistics to increase power in the case of non-proportional hazards is a well-known robust method in survival analysis, dating back to [46, 47, 48] in the 1980s. A recent review of these developments is provided by Ganju, Yu and Ma [49], who also

propose a bootstrap method, which they call "permutations", to evaluate the null distribution of the combined statistic. These works, however, assume a fixed-duration design. For time-sequential designs, the major novelty of our approach lies in (13), in which we use the "adaptive statistic" only at the prescheduled termination time t^* of the trial to attain power similar to that of a fixed-duration trial using such statistic. The modified Haybittle-Peto boundary for the time-sequential logrank test, which spends only $\varepsilon\alpha$ for the type I error in the interim analyses prior to t^*, also helps to achieve this by choosing ε suitably; see [27]. Because only a fraction ε of the type I error is spent at interim analyses, we do not want to pay for combing test statistics for combining statistics in stopping for efficacy, hence (13) only uses the logrank statistics for early stopping prior to t^*. Although we have focused on cumulative hazard or survival differences at $s_1, \ldots, s_L$ in (13), other Studentized statistics such as those in [46, 47, 48, 49] can be used instead. The choice of survival times $s_1, \ldots, s_L$ depends on domain knowledge about the disease. For example, for cardiovascular diseases, early-stage survival differences between treatment and control are usually important to consider, whereas for certain types of cancer, later-stage survival differences are important. Different choices of ψ in the censored rank statistics (1) weight the early and later failures differently. In particular, the logrank statistic gives equal weight as the proportional hazards assumption underlying the test statistic implies that the hazard ratio remains the same for the early and late events.

Early stopping for futility is discussed in [44] and Section 3.3 for the logrank or more general test statistic. For the logrank statistic, it involves predicting the total number of events by the prescheduled termination date t^*. Clearly, if too few events are predicted to occur, then the trial should either stop for futility or extend the termination date. On the other hand, the effect size $\sqrt{n}\mu(t)$ also plays an important role in the theory of futility stopping in [44] . According to this theory , evidence of futility of continuing with the logrank test only supports stopping the logrank test, but not the other tests to be carried out at time t^* that supplement the logrank test. A similar problem is considered in the next section, in which stopping for futility of testing one null hypothesis (related to patient subgroup) redirects the trial to test other relevant null hypotheses in a multiple testing framework.

4.2. Emerging trends in adaptive design of time-sequential survival trials

As noted in [37], adaptive seamless designs represent an important trend of innovations in clinical trial designs to address the need for more efficient and effective drug development processes in translating the breakthroughs in biomedical sciences into treatments of complex diseases. In particular, whereas traditional Phase III trials are inefficient "stand-alone" trials whose analyses ignore the information from previous phases, [50] argues for combining Phase II and Phase III into a single trial conducted in two stages, which is the basic idea of seamless Phase II-III designs. Although the seamless Phase II-III in Section 2.3 is one such example, what [50] focuses on for the initial stage (Phase II) is to choose a treatment regimen (such as dose) or a patient subgroup for continuation in the second stage (Phase III) of the adaptive seamless trial.

The development of imatinib, the first drug to target the genetic defects of chronic myeloid leukemia (CML) while leaving healthy cells unharmed, has revolutionized the treatment of cancer. Most new targeted treatments, however, have resulted in only modest clinical benefit, with less than 50% remission rates and less than one year of progression-free survival, unlike a few cases such as trastuzumab in HER2-positive breast cancer, imatinib in CML, and gefinitib and erlotinib in non-small cell lung cancer. While the targeted treatments are devised to attack specific targets, the "one size fits all" treatment regimens commonly used may have diminished their effectiveness and genomic-guided and risk-adapted personalized therapies that are tailored for individual patients are expected to substantially improve the effectiveness of these treatments. Of particular interest to the pharmaceutical industry is how personalized biomarker data can be used in a phase III trial for regulatory approval of a new treatment, particularly for treating cancer by attacking specific targets. There are two important preliminaries prior to designing the trial. One is to identify the biomarkers that are predictive of response, and the other is to develop a biomarker classifier that identifies patients who are sensitive to the treatment, denoted Dx+. An example is trastuzumab, for which strong evidence of the relationship between the biomarker, HER2, and the drug effect was found early and led to narrowing the patient recruitment to HER2-positive patients in the phase III trial. In the ideal setting that the biomarker classifier can partition the patient population into drug-sensitive (Dx+) and drug-resistant (Dx-) subgroups, it is clear that Dx- patients should be excluded from the clinical trial. In practice, however, the cut-point for the Dx+ group is often based on data from early

phase trials with relatively small sample sizes and has substantial statistical uncertainty (variability). Thus, a dilemma arises at the design stage of the Phase III trial. Should the trial only recruit Dx+ patients who tend to have larger effect size, or should it have broad eligibility from the entire intended-to-treat (ITT) patient population but a diluted overall treatment effect size? The former has the disadvantage of an overly stringent exclusion criterion that misses a large fraction of patients who can benefit from the treatment if the classifier imposes relatively low false positive rate for Dx+ patients, while the latter has the disadvantage of ending up with an insignificant treatment effect by including patients that do not benefit from the treatment.

Brannath et al. [51] point out the difficulties with using traditional designs to address this dilemma:

> Selecting a spurious sub-population could lead to wrongly limiting access to the treatment for only a fraction of the benefiting population. Generating the evidence to support such a development strategy traditionally requires (i) a hypothesis generating (exploratory) study to identify a sub-population, (ii) the confirmation of the sensitivity of this sub-population in an independent second(e.g. phase II) study, before (iii) running a phase III study in the selected target population. The formal claim of efficacy in the target population is to be based on the later phase III study results. Consequently, the traditional approach is very time consuming and resource intensive and does not facilitate efficient use of accumulating evidence to support the final claim of efficacy in the relevant population.

Recognizing that the adaptive seamless designs proposed in [50] can be used to "combine into a single study the objectives (ii) and (iii)", Brannath et al. [51] extend these two-stage designs to time-to-event data that may be censored at interim analysis. As in [50], they consider multiple testing of the two null hypotheses H_0 (for the ITT population) and H_{0+} (for the Dx+ sub-population), and follow the p-value combination approach introduced in [52] and [53] to combine the first-stage p-value with the second-stage p-value that is based exclusively on the second-stage data. Using the logrank statistics for time-to-event data, they extend the independent increments property of the asymptotic distribution of $S_n(t)$ in Section 2.1 to the "stratified logrank test" with different strata for the sub-population and its complement so that the ideas of [50, 52, 53] for independent normal observations can be applied. The interim analysis uses Bayesian posterior probabilities to decide whether

to continue only with Dx+ (if the posterior probability that Dx- patients can benefit from the new treatment is low), or to stop the study for futility if the posterior probabilities of benefit for ITT and for Dx+ are both low; the posterior probabilities are computed by using normal approximations to logrank statistics and Bayesian conjugate priors for normal random walks.

Jenkins, Stone and Jennison [54] have introduced a further refinement of [51] while still using the p-value combination approach. They treat ITT and Dx+ as "co-primary populations" of the two-stage trial. The interim analysis at the end of the first stage decides whether to continue with both co-primary populations, or with the Dx+ subpopulation only, or with ITT only, or to stop for futility. The decision is based on intermediate endpoint, progression-free survival (PFS) , that is correlated with the primary endpoint, overall survival, of the trial. It argues for a "simple, unequivocal" interim decision rule based on the estimated hazard ratios for PFS within the ITT and Dx+ groups. "Target values are set and the trial only continues in those groups for which the hazard ratio exceeds the target. Simulations of the clinical trial design can be used to choose the thresholds for this decision rule so as to ensure the design has higher power", and simulation studies of the design in [54] show that the proposed design does not inflate the type I error.

As noted in [37], it is widely recognized that this p-value combination approach is inefficient, and more efficient seamless Phase II-III designs incorporating patient subgroup selection are long-standing open problems. In fact, Section 4.3 of [37] and the recent paper [55] have developed such designs for comparing a new method against standard medical care for stroke patients. The endpoint of that trial is the Rankin score, which is much easier to handle than the censored failure-time endpoint in survival trials. Lai, Liao and Tsang have recently extended the approach in [55] to survival outcomes, and the method and results will be presented elsewhere. An important innovation of their work is that unlike [51] and [54] which rely heavily on the logrank statistic, a more flexible and powerful statistic of the type (13), after stratification into subgroups, is used at the prescheduled end of the trial.

References

[1] Cox DR. Regression models and life-tables. J Roy Statist Soc Ser B 1972; 34:187–220.

[2] Cox DR. Partial likelihood. Biometrika 1975; 62(2):269–276.

[3] Andersen PK, Gill RD. Cox's regression model for counting processes: A large sample study. Ann Stat 1982; 10:1100–1120.

[4] Begun JM, Hall WJ, Huang WM, Wellner JA. Information and asymptotic efficiency in parametric-nonparametric models. Ann Stat 1983; 11:432–452.

[5] Andersen PK, Borgan Ø, Gill RD, Keiding N. Statistical Models Based on Counting Processes. New York: Springer; 1993.

[6] Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer; 2003.

[7] Beta-Blocker Heart Attack Trial Research Group. A randomized trial of propranolol in patients with acute myocardial infarction: 1. mortality results. J Am Med Assoc 1982; 247(12):1707–1714.

[8] Beta-Blocker Heart Attack Trial Research Group. Beta-blocker heart attack trial: design, methods, and baseline results. Contemp Clin Trials 1984; 5(4):382–437.

[9] Gu M, Lai TL. Determination of power and sample size in the design of clinical trials with failure-time endpoints and interim analyses. Contemp Clin Trials 1999; 20:423–438.

[10] Lan KKG, Demets DL. Group sequential procedures: Calendar versus information time. Stat Med 1989; 8:1191–1198.

[11] Chuang CS, Lai TL. Hybrid resampling methods for confidence intervals (with discussion and rejoinder). Statist. Sinica 2000; 10:1–50.

[12] Lai TL, Li W. Confidence intervals in group sequential trials with random group sizes and applications to survival analysis. Biometrika 2006; 93:641–654.

[13] Lai TZ, Shih MC, Su Z. Tests and confidence intervals for secondary endpoints in sequential clinical trials. Biometrika 2009; 96(4):903–915.

[14] Gu M, Lai TL. Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials. Ann Statist 1991; 19(3):1403–1433.

[15] Gu M, Lai TL. Repeated significance testing with censored rank statistics in interim analysis of clinical trials. Statist Sinica 1998; 8(2):411–428.

[16] Self SG. An adaptive weighted logrank test with application to cancer prevention and screening trials. Biometrics 1991; 47:975-986.

[17] Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. Biometrika 1982; 69(3):553–566.

[18] Peto R, Peto J. Asymptotically efficient rank invariant test procedures. J Roy Statist Soc Ser A 1972; 135(2):185–207.

[19] Prentice RL. (1978) Linear rank tests with right censored data. Biometrika 1978; 65(1):167–179.

[20] Slud E, Wei LJ. (1982) Two-sample repeated significance tests based on the modified Wilcoxon statistic. J Am Statist Assoc 1982; 77(380):862–868.

[21] Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. J Roy Statist Soc Ser A 1969; 132:235–244.

[22] Lan KKG, Demets DL. Discrete sequential boundaries for clinical trials Biometrika 1983; 70 (3):659–663.

[23] Scharfstein DO, Tsiatis AA. The use of simulation and bootstrap in information-based group sequential studies. Stat Med 1998; 17(1):75–87.

[24] Scharfstein DO, Tsiatis AA, Robins JM. Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. J Am Statist Assoc 1997; 92(440):1342–1350.

[25] Haybittle J. Repeated assessment of results in clinical trials of cancer treatment. Brit J Radiol 1971; 44(526):793797.

[26] Peto R, Pike M, Armitage P, Breslow N, Cox D, Howard S, Mantel N, McPherson K, Peto J, Smith P. Design and analysis of randomized clinical trials requiring prolonged observation of each patient 1. Introduction and design. Br J Cancer 1976; 34(6):585–612.

[27] Lai TL, Shih MC. (2004) Power, sample size and adaptation considerations in the design of group sequential clinical trials. Biometrika 2004; 91(3):507–528.

[28] Siegmund D. Estimation following sequential tests. Biometrika 1978; 65:341–49.

[29] Lan KKG, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. Comm Statist CSequent Anal 1982; 1(3):207–219.

[30] Lin DY, Yao Q, Ying Z. A general theory on stochastic curtailment for censored survival data J Am STat Assoc 1999; 94:510–521.

[31] Spiegelhalter DJ, Freedman LS, Blackburn DR. Monitoring clinical trials: Conditional or predictive power? Contemp Clin Trials 1986; 7:8–17.

[32] Ibrahim JG, Chen M-H, Sinha D. Bayesian Survival Analysis. New York:Springer; 2001.

[33] Ishwaran H, James L. Computational methods for multiplicative intensity models using weighted gamma processes: Proportional hazards, marked point processes, and panel count data. J Am Stat Assoc 2004; 99:175–190.

[34] Hobbs BP, Carlin BP. Practical Bayesian design and analysis for drug and device trials. J Biopharm Stat 2008; 18:54–80.

[35] Berry SM, Carlin BP, Lee JJ, Müller P. Bayesian Adaptive Methods for Clinical Trials. Boca Ration FL: Chapman & Hall/CRC; 2011.

[36] Berry DA. Bayesian clinical trials. Nature Rev Drug Disc 2006; 5:27–36.

[37] Lai TL, Lavori PW, Tsang KW. Adaptive design of confirmatory trials: Advances and challenges. Contemp Clin Trials 2015; this issue.

[38] Simon R. Optimal 2-stage designs for phase II clinical trials. Contemp Clin Trials 1989; 10(1):1–10.

[39] Vickers AJ, Ballen V, Scher HI. Setting the bar in phase II trials: the use of historical data for determining "go/no go" decision for definitive phase III testing. Clin Cancer Res 2007; 13(3):972–976.

[40] Rubinstein L, Crowley J, Ivy P, LeBlanc M, Sargent D. Randomized phase II designs. Clin Cancer Res 2009; 15:1883–1890.

[41] Inoue LYT, Berry DA, Thall PF. Seamlessly expanding a randomized Phase II trial to Phase III. Biometrics 2002; 58:823–831.

[42] Huang X, Ning J, Li Y, Estey E, Issa JP, Berry DA. Using short-term response information to facilitate adaptive randomization for survival clinical trials. Stat Med 2009; 28(12):1680–1689.

[43] Lai TL, Lavori PW, Shih MC. Sequential design of phase IIIII cancer trials. Stat Med 2012a; 31(18):1944–1960.

[44] He P, Lai TL, Liao OY-W. Futility stopping in clinical trials. Stat & Its Interface 2012; 5:415–423.

[45] Susarla V, Van Ryzin J. Nonparametric Bayesian estimation of survival curves from incomplete observations. J Am Stat Assoc 1976; 71:897–902.

[46] Tarone RE. On the distribution of the maximum of the logrank statistic and the modified Wilcoxon statistic. Biometrics 1981; 37:79–85.

[47] Gastwirth JL. The use of maximin efficiency robust tests in combining contingency tables and survival analysis. JASA 1985; 80:380–384.

[48] Fleming TR, Harrington DP, O'sullivan M. Supremum versions of the log-rank and generalized Wilcoxon statistics. JASA 1987; 82:312–320.

[49] Ganju J, Yu X, Ma G. Robust inference from multiple test statistics via permutations: a better alternative to the single test statistic approach for randomized trials. Pharm Stat 2013; 12:282–290.

[50] Bretz F, Schmidli H, König F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. Biom J 2006; 4:623–634.

[51] Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M, Racine-Poon A. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. Stat Med 2009; 28:1445–1463.

[52] Bauer P. Multistage testing with adaptive designs (with Discussion). Biom Inform Med Bio 1989; 20:130–148.

[53] Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses Biometrics 1994; 50:1029-1041.

[54] Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. Pharm Stat 2011; 10:347-356.

[55] Lai TL, Lavori PW, Liao OY. Adaptive choice of patient subgroup for comparing two treatments. Contemp Clin Trials 2014; 39:191–200.