Statistica Sinica **17**(2007), 1567-1589

DESIGNED EXTENSION OF SURVIVAL STUDIES: APPLICATION TO CLINICAL TRIALS WITH UNRECOGNIZED HETEROGENEITY

Yi Li[1], Mei-Chiung Shih[2] and Rebecca A. Betensky[1]

[1]*Harvard University and* [2]*Standford University*

Abstract: It is well known that unrecognized heterogeneity among patients, such as is conferred by genetic subtype, can undermine the power of a randomized trial, designed under the assumption of homogeneity, to detect a truly beneficial treatment. We consider the conditional power approach to allow for recovery of power under unexplained heterogeneity. While Proschan and Hunsberger (1995) confined the application of conditional power design to normally distributed observations, we consider more general and difficult settings in which the data are in the framework of continuous time and are subject to censoring. In particular, we derive a procedure appropriate for the analysis of the weighted log rank test under the assumption of a proportional hazards frailty model. The proposed method is illustrated through application to a brain tumor trial.

Key words and phrases: Adaptive design, conditional power, frailty model.

1. Introduction

It is well known that unrecognized heterogeneity among patients, such as is conferred by genetic subtype, can undermine the power of a randomized trial to detect a truly beneficial treatment (Betensky, Louis and Cairncross (2002), Li, Betensky, Louis and Cairncross (2002)). One mechanism through which genetic subtype might affect study power is that of precision. This operates if a treatment is equally effective in all genetic subtypes, but genetic subtype is itself predictive of survival and not associated with treatment. The North American and European Intergroup trials comparing chemotherapy plus radiotherapy versus radiotherapy alone for patients with anaplastic oligodendroglioma, a type of malignant brain tumor, were designed prior to the discovery of at least three clinically distinct genetic subtypes among patients with the histological diagnosis of this disease (Ino, Betensky, Zlatescu, Sasaki, Macdonald, Stemmer-Rachamimov, Ramsay, Cairncross and Louis (2001)). Patients with allelic loss of chromosome 1p have long survival times. In contrast, patients with chromosome 1p intact and no mutation of the TP53 gene have short survival times. Patients with chromosome 1p intact and a TP53 mutation follow an intermediate course. As these

genetic subtypes were revealed, the investigators questioned whether the effect of chemotherapy versus radiotherapy would be attenuated due to failure to adjust for the "precision variable" of genetic subtype.

This scenario illustrates that until all of the genetic underpinings of disease are discovered, clinical trials will likely suffer from insufficient power due to the as yet unrecognized heterogeneity of disease. One partial solution to this problem is to collect specimens from all study participants for retrospective analysis once the genetics of the disease are understood. Although this will lead to added precision in the analyses, there may not be enough subjects of certain genetic subtypes for adequately powered analyses. Further, this discovery is unlikely to occur contemporaneously with the completion of the clinical trial. Also, in many diseases, unknown environmental influences likewise contribute to heterogeneity and there is even less direction in the search for these influences.

Another solution is to design clinical trials in recognition of this problem. In the absence of genetic information to explain subject heterogeneity, a new treatment can be proven effective only if its average effect among all patients exceeds the average effect of the standard treatment. Typically, the marginalized hazard function for new treatment will not be proportional to the marginalized hazard for the standard treatment. This is true, for example, if the hazard function for the new treatment, conditional on an unobserved individual frailty, is proportional to that for the standard treatment. Another complicating feature is that pilot data may not be reflective of population-based data due to inadvertent selection biases (Betensky, Cairncross and Louis (2003)). For these reasons, a trial design based on comparing average treatment effects via the log rank test, with parameters estimated from pilot data, is likely to be underpowered (Betensky, Louis and Cairncross (2002)). To recover this lost power in the absence of genetic information, we consider a conditional power design in the setting of the weighted log rank test.

Proschan and Hunsberger (1995) proposed a two-stage design that uses information about the significance of the treatment effect after the first stage to determine the number of additional subjects required, and the critical value to use after the second stage. In contrast to most adaptive designs, this procedure uses the first stage to estimate the treatment effect, rather than the nuisance parameters. Without proper control, the actual type I error rate can be more than double the originally planned rate. Proschan and Hunsberger (1995) described how to control this error rate.

While Proschan and Hunsberger (1995) confined their application to normally distributed observations, adaptive designs have recently aroused much research interest for censored observations. For example, Schaefer and Mueller (2001) and Shen and Cai (2003) have considered sample size re-estimation for

multiple-stage designs with censored outcomes. However, virtually no work has been done to accommodate heterogeneities detected at the interim time points. This paper derives a procedure appropriate for the analysis of the weighted log rank test under the assumption of a proportional hazards frailty model, where the frailties are introduced to model unobserved heterogeneities. We consider separately two cases of potential practical interest in the setting of survival data: extension of follow-up time without additional subject accrual, and extension of follow-up time in conjunction with additional accrual. The approach selected in practice will depend on the relative costs of additional follow-up time versus additional subjects. It will depend also on the relative amounts of additional information expected through each means of extension. That is, if the treatment comparison at the end of the first stage is not significant, but there is little additional information expected within any amount of additional follow-up, additional subjects will be required. Although our investigations of these designs were motivated by the need to recover power in the presence of unrecognized heterogeneity within a proportional hazards frailty model, many of our results hold more generally for any analysis based on the log rank test.

In Section 2 we present notation and describe the frailty models for unrecognized heterogeneity. In Section 3, we derive the conditional power for the weighted log rank test, with allowance for added follow-up time after the first stage of the study. In Section 4, we describe the details of the conditional power procedure for added follow-up time. We investigate the unconditional power of the conditional power procedure in Section 5. In Section 6, we derive the conditional power for the weighted log rank test, with allowance for both added follow-up and added subjects after the first stage of the study. We describe the details of this conditional power procedure in Section 7. We discuss application of the procedures to the frailty model in Section 8, and we apply the procedure to a brain tumor study in Section 9. Section 10 contains a discussion of these designs.

2. Frailty Proportional Hazards Model

Suppose each of n subjects, who may enter the study in a 'staggered entry' fashion, is randomly assigned to treatment 1 with probability a_1 and to treatment 2 with probability a_2, leading to n_1 subjects receiving treatment 1 and n_2 subjects receiving treatment 2. Let Z be the treatment indicator, with $Z = 0$ for subjects who are assigned treatment 1 and $Z = 1$ for treatment 2. Associated with each subject is a survival time (measured from study entry), T_i, and a censoring time, C_i, distributed according to G, such that $G(c) = P(C > c)$. Only $X_i = \min(T_i, C_i)$ is observed, along with an indicator, δ_i, for whether death occurred: $\delta_i = 1$ if $T_i \leq C_i$ and $\delta_i = 0$ if $T_i > C_i$. Also associated with each

subject is an unobserved frailty, b_i, where $b_i \sim F(b;\theta)$ and θ parameterizes the frailty distribution, F. The frailty captures the unexplained heterogeneity among subjects (e.g., Oakes (1989)).

We assume that conditional on the unobserved frailty, b, the survival time, T, follows a proportional hazards model

$$P(t \le T < t + dt | T \ge t, Z, b) = \lambda_0(t) \exp(\beta Z + b) dt. \tag{1}$$

Here, β measures the main effect for treatment. We also assume that this conditional proportional hazards model is valid throughout the study period; similar proportionality assumptions have been made in other conditional power literatures (see, e.g., Betensky (1998)). In general, the unconditional treatment-specific hazard functions induced from (1) do not follow a proportional hazards model (Betensky, Louis and Cairncross (2002)). An exception is the positive stable frailty model, which preserves the marginal proportionality (Hougaard (1986)).

Under (1), the marginal cumulative hazard function for the $Z = 1$ group is $\Lambda(t|\beta) = -\log(P\{\Lambda_\beta(t)\})$ and the hazard function is

$$\lambda(t|\beta) = -\frac{d}{dt}\log(P\{\Lambda_\beta(t)\}) = \lambda_0(t)\exp(\beta)\frac{-P'\{\Lambda_\beta(t)\}}{P\{\Lambda_\beta(t)\}},$$

where $\Lambda_\beta(t) = \Lambda_0(t)\exp(\beta)$, $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$, and $P(s) = \int e^{-s\exp(b)} dF(b;\theta)$ is the Laplace transform for the random variable $\exp(b)$. The marginal cumulative hazard function and hazard function for the $Z = 0$ group are denoted by $\Lambda(t|0)$ and $\lambda(t|0)$, respectively. Assuming differentiability of $\lambda(t|\beta)$ with respect to β in a neighborhood of 0, a Taylor series expansion yields

$$\lambda(t|\beta) \doteq \lambda(t|0) + \beta \frac{\partial}{\partial\beta}\bigg|_{\beta=0} \lambda(t|\beta).$$

While in our motivating brain tumor example we considered the case of three genetic subtypes, in reality we expect subjects' complete genetic profiles to confer continua of risk. Thus, we consider the frailty to be a continuous random variable. When the frailty, b, follows certain common distributions, closed-form expressions for the Laplace transform function $P(s)$ are available:

1. normal frailty: $P(s) = \int_{-\infty}^{\infty} \exp(-se^{\sqrt{\theta}x})\phi(x)dx$;
2. log gamma frailty: $P(s) = (1 + \theta s)^{-1/\theta}$;
3. inverse Gaussian frailty (Hougaard (1984)) : $P(s) = \exp\{-\theta[(1+2\theta s)^{1/2} - 1]\}$;
4. positive stable (Hougaard (1986)): $P(s) = \exp(-s^\theta)$.

In each of these, $\theta > 0$ is the variance of the frailty. When the frailty follows one of these distributions, $\lambda(t|\beta)$ increases with β for fixed t.

3. Weighted Log Rank Test and Conditional Power for Adding Follow-up

In this section, we compute the conditional power for the weighted log rank test. Many components of these results, and those of Sections 4-7, are completely general and are not confined to any specific model, such as the frailty models of Section 2. In Sections 8 and 9 we discuss application of this procedure to data that follow the frailty models of Section 2.

The log rank test can be written in the following general form:

$$L(t) = \int_0^t K(u)\frac{d\bar{N}_1(u)}{\bar{Y}_1(u)} - \int_0^t K(u)\frac{d\bar{N}_2(u)}{\bar{Y}_2(u)}, \tag{2}$$

where, for $k = 1, 2$, the total number of subjects at risk in each treatment group at time t is $\bar{Y}_k(t) = \sum_{i=1}^n Y_i(t)I(Z_i = k-1)$, the total number of observed failures in each treatment group by t is $\bar{N}_k(t) = \sum_{i=1}^n N_i(t)I(Z_i = k-1)$, and

$$K(s) = w(s)\left(\frac{n_1+n_2}{n_1 n_2}\right)^{\frac{1}{2}} \frac{\bar{Y}_1\bar{Y}_2}{\bar{Y}_1+\bar{Y}_2},$$

where $w(s)$ is a predicatable weight function. Commonly used weight functions can be found in Fleming and Harrington (1991, p.257). In particular, $w(s) \equiv 1$ corresponds to the simple log rank test.

Under a sequence of local alternatives for β that converges to zero at the appropriate rate as the sample size increases to infinity, the log rank statistic has asymptotically a finite mean and variance. In particular, we have the following asymptotic result.

Proposition 1. *Assume that for each $k = 1, 2$, there exists a function π_k and a constant a_k such that*

$$\sup_{t\geq 0}\left|\frac{\bar{Y}_k(t)}{n_k} - \pi_k(t)\right| \xrightarrow{p} 0, \tag{3}$$

$$\frac{n_k}{n} \to a_k. \tag{4}$$

Assume further that, under a sequence of local alternatives converging to null at the rate of $n^{-1/2}$, say, $\beta_n = [(n_1+n_2)/(n_1n_2)]^{1/2}\beta_0$,

$$\sup_{t>0}\left|\Lambda(t|\beta_n) - \Lambda(t|0)\right| \to 0. \tag{5}$$

Then, as $n \to \infty$, on any finite interval, $[0, \tau]$, where $\tau < \tau_0 = \sup\{t > 0 : \pi_1(t)\pi_2(t) > 0\}$, the weighted log rank test $L(\cdot)$ converges weakly to a tight Gaussian process $L^\infty(\cdot)$, which takes form $L^\infty(t) = \mu(t) + B\{v(t)\}$, where $B\{t\}$ is a

standard Brownian motion, drift $\mu(t) = \int_0^t \kappa(u)\gamma(u)d\Lambda(u|\beta_0)$, *and variance function* $v(t) = \int_0^t [(a_1\pi_1(u) + a_2\pi_2(u))/(\pi_1(u)\pi_2(u))]\kappa^2(u)d\Lambda(u|\beta_0)$. *Here* $\kappa(t)$ *is the probabilistic limit of* $\{(n_1+n_2)/(n_1n_2)\}^{1/2}K(t)$, *and* $\gamma(t)$ *is the probabilistic limit of* $\{(n_1n_2)/(n_1+n_2)\}^{1/2}\{[\lambda(t|\beta_0)/\lambda(t|0)] - 1\}$.

This proposition is a direct application of Theorem 7.4.1 in Fleming and Harrington (1991, p.269) after verifying that their conditions (4.5)-(4.8) hold on any finite interval $[0, \tau]$. From this proposition, it follows that $L^\infty(t)$ has mean $\mu(t)$, variance $v(t)$, and covariance $Cov\{L^\infty(t), L^\infty(s)\} = v\{\min(s,t)\}$. Some algebra shows that

$$\kappa(u) = \frac{w(u)\pi_1(u)\pi_2(u)}{a_1\pi_1(u) + a_2\pi_2(u)},$$

and, hence,

$$v(t) = \int_0^t w^2(u)\frac{\pi_1(u)\pi_2(u)}{a_1\pi_1(u) + a_2\pi_2(u)}d\Lambda(u|0) = \int_0^t w(u)\kappa(u)d\Lambda(u|0).$$

Note that when $\beta > 0$, $\gamma(u) > 0$ since $\lambda(t|\beta) > \lambda(t|0)$ under the frailty model. This implies that $\mu(t)$ increases with t when $\beta > 0$. This, of course, is not the case for all applications of the weighted log rank test, but is particular to the frailty model (1).

At a pre-determined time t_1 (e.g., interim analysis time), the normalized test statistic is given by $l_{t_1} = L(t_1)/\hat{v}^{1/2}(t_1)$, where $\hat{v}(\cdot)$ is obtained from $v(\cdot)$ by replacing θ with $\hat{\theta}$. It follows from Proposition 1 that $CP_{\beta_0}(t, z_\alpha|l_{t_1})$, the conditional probability that the normalized test statistic at time t, $L(t)/\hat{v}^{1/2}(t)$, will exceed z_α, given the value of the statistic at time t_1 and given that $\beta_n = (n_1 + n_2/n_1n_2)^{1/2}\beta_0$ is

$$\begin{aligned}
CP_{\beta_0}(t, z_\alpha|l_{t_1}) &= P\left(\frac{L(t)}{\hat{v}^{\frac{1}{2}}(t)} > z_\alpha \middle| \frac{L(t_1)}{\hat{v}^{\frac{1}{2}}(t_1)} = l_{t_1}, \beta_0\right) \\
&\doteq P\left(\frac{L^\infty(t)}{v^{\frac{1}{2}}(t)} > z_\alpha \middle| \frac{L^\infty(t_1)}{v^{\frac{1}{2}}(t_1)} = l_{t_1}, \beta_0\right) \\
&= P\left(\frac{L^\infty(t) - L^\infty(t_1) + L^\infty(t_1)}{v^{\frac{1}{2}}(t)} > z_\alpha \middle| \frac{L^\infty(t_1)}{v^{\frac{1}{2}}(t_1)} = l_{t_1}, \beta_0\right) \\
&= P\left(L^\infty(t) - L^\infty(t_1) > z_\alpha v^{\frac{1}{2}}(t) - l_{t_1}v^{\frac{1}{2}}(t_1) \middle| \beta_0\right) \\
&= P\left(\frac{L^\infty(t) - L^\infty(t_1) - \{\mu(t) - \mu(t_1)\}}{\{v(t) - v(t_1)\}^{\frac{1}{2}}}\right. \\
&\quad \left. > \frac{z_\alpha v^{\frac{1}{2}}(t) - l_{t_1}v^{\frac{1}{2}}(t_1) - \{\mu(t) - \mu(t_1)\}}{\{v(t) - v(t_1)\}^{\frac{1}{2}}} \middle| \beta_0\right).
\end{aligned}$$

The $\doteq$ above stems from the weak convergence of $L(\cdot)$ to $L^{\infty}(\cdot)$ and the last two equalities follow from the independent increment property for a Brownian motion. Hence,

$$CP_{\beta_0}(t, z_\alpha | l_{t_1}) \doteq 1 - \Phi\left(\frac{z_\alpha v^{\frac{1}{2}}(t) - l_{t_1} v^{\frac{1}{2}}(t_1) - [\mu(t) - \mu(t_1)]}{[v(t) - v(t_1)]^{\frac{1}{2}}}\right), \tag{6}$$

where $\Phi(\cdot)$ is the cdf for a standard normal distribution.

4. Designed Extension via Conditional Power: Adding Follow-up

In this section, we propose an extension of the two-stage design of Proschan and Hunsberger (1995) to the weighted log rank test, with possible extension of follow-up time. The key idea behind this methodology is based on the data of the first stage, to determine the type I error level to use at the second stage, and then apply a logrank test at this level to events occurring in the second stage, while maintaining the overall type I error at a prescribed level (say, 0.05). Specifically, at the time of the interim analysis, t_1, we calculate the additional follow-up time, $t - t_1$, and the critical value, z_α, such that the conditional power, $CP_{\beta_0}(t, z_\alpha | l_{t_1}) = 1 - \delta$. The type I error rate for this design is $\int_{-\infty}^{\infty} CP_0(t, z_\alpha | l_{t_1}) \phi(l_{t_1}) dl_{t_1}$ where $\phi(\cdot)$ is the standard normal density.

If t is chosen without regard to controlling the overall type I error rate, it can far exceed the nominal α level. To see this, consider choosing t to maximize the null conditional power

$$CP_0(t, z_\alpha | l_{t_1}) = P_0\left(\frac{L(t)}{\sqrt{v(t)}} > z_\alpha \middle| \frac{L(t_1)}{\sqrt{v(t_1)}} = l_{t_1}\right) = 1 - \Phi\left(\frac{z_\alpha \sqrt{v(t)} - l_{t_1}\sqrt{v(t_1)}}{\sqrt{v(t) - v(t_1)}}\right).$$

When $l_{t_1} > z_\alpha$, $t = t_1$ is the maximizer, with $CP_0 = 1$. Otherwise $CP_0(t, z_\alpha | l_{t_1}) = 1 - \Phi((z_\alpha\sqrt{1+R} - l_{t_1})/\sqrt{R})$, where $R = v(t)/v(t_1) - 1$ is the additional information (expected events) when the study is extended to time t, expressed as multiples of the information available at interim t_1. Denote by $R_{\max} = v(\infty)/v(t_1) - 1$ the maximum additional information that could be added. Following Proschan and Hunsberger (1995), we obtain the maximum type I error rate

$$\begin{aligned}
\alpha_{\max} = &\int_{-\infty}^{0} \left[1 - \Phi\left(\frac{z_\alpha\sqrt{1+R_{\max}} - l_{t_1}}{\sqrt{R_{\max}}}\right)\right] \phi(l_{t_1}) dl_{t_1} \\
&+ \int_{0}^{\frac{z_\alpha}{\sqrt{1+R_{\max}}}} \left[1 - \Phi\left(\frac{z_\alpha\sqrt{1+R_{\max}} - l_{t_1}}{\sqrt{R_{\max}}}\right)\right] \phi(l_{t_1}) dl_{t_1} \\
&+ \int_{\frac{z_\alpha}{\sqrt{1+R_{\max}}}}^{z_\alpha} \left[1 - \Phi\left(\sqrt{z_\alpha^2 - l_{t_1}^2}\right)\right] \phi(l_{t_1}) dl_{t_1} + \int_{z_\alpha}^{\infty} \phi(l_{t_1}) dl_{t_1}.
\end{aligned} \tag{7}$$

When $R_{\max} = \infty$, (7) reduces to (A1) of Proschan and Hunsberger (1995). For significance level $\alpha = 0.05$ with $z_\alpha = 1.645$, Table 1 lists the maximum type I error, for a range of $R_{\max}$.

Table 1. Maximum type I error.

$R_{\max}$	0.0	0.05	0.1	0.5	1	2	5	10	∞
$\alpha_{\max}$	0.050	0.059	0.063	0.075	0.082	0.089	0.097	0.102	0.115

This inflation of the type I error can be avoided by using a larger critical value $k > z_\alpha$ (at stages 1 and 2) and not continuing the study unless the p-value at time t_1 is less than some number p^*, where $p^* \le 0.5$. In this case, $\alpha_{\max} = \int_{-\infty}^{\infty} A_0(l_{t_1})\phi(l_{t_1})\, dl_{t_1}$, where

$$A_0(l_{t_1}) = \begin{cases} 0 & \text{if } l_{t_1} < z_{p^*} \\ 1 - \Phi\Big(\frac{k\sqrt{1+R_{\max}} - l_{t_1}}{\sqrt{R_{\max}}}\Big) & \text{if } z_{p^*} \le l_{t_1} < \max(z_{p^*}, \frac{k}{\sqrt{1+R_{\max}}}) \\ 1 - \Phi\Big(\sqrt{k^2 - l_{t_1}^2}\Big) & \text{if } \max(z_{p^*}, \frac{k}{\sqrt{1+R_{\max}}}) \le l_{t_1} < k \\ 1 & \text{if } l_{t_1} \ge k. \end{cases}$$

For a given $R_{\max} = v(\infty)/v(t_1) - 1$, we choose p^* and k such that $\int_{-\infty}^{\infty} A_0(z_1)\phi(z_1)\, dz_1 = \alpha$. This conditional error function contains both "circular" and "linear" components (Proschan and Hunsberger (1995)). Table 2 lists the values of k corresponding to different choices of p^* and $R_{\max}$ for $\alpha = 0.05$. As expected, k increases with respect to both p^* and $R_{\max}$.

Table 2. Adjusted critical value, k, as a function of p^* and $R_{\max}$.

	p^*						
$R_{\max}$	0.1	0.15	0.20	0.25	0.30	0.40	0.50
0.1	1.745	1.754	1.756	1.756	1.756	1.756	1.756
0.5	1.772	1.812	1.829	1.838	1.843	1.847	1.848
1.0	1.773	1.820	1.846	1.862	1.872	1.884	1.889
2.0	1.773	1.821	1.851	1.872	1.888	1.908	1.920
5.0	1.773	1.821	1.852	1.875	1.894	1.921	1.941
10.0	1.773	1.821	1.852	1.875	1.894	1.924	1.947
∞	1.773	1.821	1.852	1.875	1.894	1.925	1.951

This maximal error function is one example of an increasing function with range $[0, 1]$, $A(l_{t_1})$, that satisfies the type I error requirement, $\int_{-\infty}^{\infty} A(l_{t_1})\phi(l_{t_1})dl_{t_1}$

$= \alpha$. The general conditional power procedure proposed by Proschan and Hunsberger (1995) can be defined with respect to any such function, $A(l_{t_1})$. For example, one advantage to a linear conditional error function is that if extension of the trial is unnecessary, the test at the end is the same as it would be for a fixed sample test. Now we consider the choice of t based on any such function. Setting $CP_0(t, c(t)|l_{t_1}) = A(l_{t_1})$ and solving for $c(t)$ yields

$$c(t) = \frac{\sqrt{v(t_1)}l_{t_1} + \sqrt{v(t) - v(t_1)}z_A}{\sqrt{v(t)}} = \frac{l_{t_1} + \sqrt{R}z_A}{\sqrt{1+R}}, \tag{8}$$

where $z_A = z_{A(l_{t_1})}$ and $R = v(t)/v(t_1) - 1$. Note that however we choose t, if the critical point c is chosen to satisfy (8), then we have an α-level procedure. In particular, we may choose t such that the conditional power under the alternative is large, e.g., ($CP_\beta(t, c|z_1) = 1-\delta$.) Plugging (8) in the expression for $CP_\beta(t, c|z_1)$ yields

$$CP_\beta(t, c|z_1) = 1 - \Phi\left(z_A - \frac{\mu(t) - \mu(t_1)}{\sqrt{v(t) - v(t_1)}}\right) = 1 - \delta.$$

Therefore, it is possible to achieve conditional power $1-\delta$ at time t if there exists a value t such that

$$z_A + z_\delta = \frac{\mu(t) - \mu(t_1)}{\sqrt{v(t) - v(t_1)}}. \tag{9}$$

As noted in Section 2, a solution does exist in the context of proportional hazards frailty models with many common frailty distributions, as in these cases $\mu(t)$ is increasing in t.

5. Comparison to Single Stage Test

Although the two-stage procedure is designed to achieve a certain level of conditional power, in many applications *unconditional* power is important as well. We conducted a simulation study to investigate the unconditional power properties of this design. Specifically, we assumed the proportional hazards log gamma frailty model (1) for $n = 100$ subjects, assigned with equal probability to each treatment, with $\lambda_0(t) = 1$, $\beta = 0.8$, and with frailty distributed according to a log gamma, Gaussian, or inverse Gaussian distribution with variance $\theta = 0.3$ or 0.8. We examined $R_{\max} = 0.50, 1.5, 5.0$, conditional power threshold $1 - \delta = 0.5, 0.8$, and we took $p^* = 0.15$. For each parameter configuration and $R_{\max}$, we calculated the time of the first analysis, t_1, to satisfy $v(\infty)/v(t_1) - 1 = R_{\max}$, and the time for the final analysis, t, to achieve conditional power $1 - \delta$. Presented in Tables 3 and 4 are the type I error rates for this two-stage procedure and

the unconditional power at times t_1 and t, estimated based on 5,000 repetitions. Also given in Table 4 is the proportional increase in follow-up time required to achieve the given conditional power. For comparison, all of these quantities are also given in Table 4 when the no-frailty model is used for analysis. We note, though, among the scenarios we have examined, the additional gains in power by including the frailty over the no-frailty model seemed to be limited.

Table 3 illustrates that the type I error is preserved under misspecification of the frailty distribution. It is not surprising that there is increased power under the two-stage design, as seen in Table 4, under correct specification of the log gamma frailty distribution. What is of interest is the varying costs, in terms of additional follow-up time, of the increased power. When the power of the fixed sample test is high, as it is at $\theta = 0.3$ and $R_{\max} = 0.5$, the follow-up time is increased by less than 5% to achieve the specified conditional power thresholds. In contrast, when the power of the fixed sample test is low, as it is at $\theta = 0.8$ and $R_{\max} = 5.0$, a 50% increase in follow-up time accompanies an increase in power from 30% to 50%. Also of note is the limitation of this procedure; by adding follow-up time without adding subjects, the unconditional power may be bounded well below one. The results in Table 4 suggest that the assumption of the log gamma frailty model is robust to misspecification, as the unconditional power remains high even when the frailty is truly normal or inverse Gaussian. However, the omission of the frailty from the model leads to a small, but consistent, decrease in unconditional power.

Table 3. Type I error rates under possible frailty misspecification as log gamma.

			true frailty distribution		
θ	$R_{\max}$	$1-\delta^{\dagger}$	log gamma	normal	inverse Gaussian
0.3	0.5	0.50	0.050	0.051	0.053
		0.80	0.051	0.057	0.049
	1.5	0.50	0.049	0.051	0.045
		0.80	0.053	0.047	0.047
	5.0	0.50	0.046	0.045	0.036
		0.80	0.045	0.038	0.041
0.8	0.5	0.50	0.048	0.045	0.051
		0.80	0.048	0.052	0.056
	1.5	0.50	0.049	0.049	0.047
		0.80	0.044	0.050	0.051
	5.0	0.50	0.046	0.045	0.040
		0.80	0.043	0.045	0.042

† Conditional power threshold for the two-stage procedure.

Table 4. Simulated unconditional power of conditional power procedure.

true model	θ	$R_{\max}$	$1-\delta^{\dagger}$	log gamma frailty analysis power at t_1	log gamma frailty analysis power at t	log gamma frailty analysis $\frac{t-t_1}{t_1}$	no frailty analysis power at t_1	no frailty analysis power at t	no frailty analysis $\frac{t-t_1}{t_1}$
log gamma	0.3	0.5	0.5	0.837	0.873	0.011	0.817	0.858	0.023
			0.8	0.835	0.894	0.039	0.829	0.893	0.112
		1.5	0.5	0.736	0.790	0.037	0.709	0.773	0.053
			0.8	0.727	0.818	0.112	0.711	0.801	0.201
		5.0	0.5	0.435	0.521	0.216	0.432	0.518	0.238
			0.8	0.454	0.577	0.577	0.427	0.563	0.720
	0.8	0.5	0.5	0.647	0.698	0.010	0.628	0.695	0.051
			0.8	0.649	0.726	0.029	0.628	0.735	0.237
		1.5	0.5	0.584	0.649	0.038	0.566	0.638	0.082
			0.8	0.591	0.695	0.103	0.566	0.679	0.290
		5.0	0.5	0.398	0.480	0.177	0.354	0.444	0.263
			0.8	0.377	0.504	0.467	0.359	0.498	0.773
normal	0.3	0.5	0.5	0.858	0.893	0.009	0.848	0.886	0.018
			0.8	0.864	0.913	0.032	0.842	0.902	0.098
		1.5	0.5	0.762	0.814	0.032	0.754	0.805	0.049
			0.8	0.756	0.836	0.103	0.740	0.832	0.166
		5.0	0.5	0.470	0.555	0.210	0.452	0.538	0.235
			0.8	0.452	0.586	0.568	0.468	0.604	0.682
	0.8	0.5	0.5	0.718	0.768	0.008	0.711	0.762	0.036
			0.8	0.736	0.799	0.024	0.711	0.802	0.183
		1.5	0.5	0.664	0.719	0.029	0.634	0.700	0.067
			0.8	0.650	0.744	0.088	0.616	0.733	0.247
		5.0	0.5	0.460	0.546	0.167	0.440	0.527	0.242
			0.8	0.451	0.585	0.446	0.435	0.568	0.742
inverse Gaussian	0.3	0.5	0.5	0.735	0.794	0.018	0.705	0.772	0.037
			0.8	0.727	0.820	0.063	0.706	0.805	0.184
		1.5	0.5	0.582	0.651	0.059	0.541	0.620	0.088
			0.8	0.556	0.678	0.182	0.542	0.662	0.294
		5.0	0.5	0.285	0.369	0.262	0.279	0.362	0.296
			0.8	0.293	0.424	0.659	0.276	0.414	0.884
	0.8	0.5	0.5	0.738	0.787	0.007	0.684	0.745	0.044
			0.8	0.735	0.799	0.023	0.696	0.797	0.192
		1.5	0.5	0.622	0.686	0.034	0.568	0.642	0.079
			0.8	0.617	0.714	0.098	0.572	0.688	0.271
		5.0	0.5	0.356	0.440	0.184	0.335	0.423	0.269
			0.8	0.359	0.494	0.485	0.362	0.492	0.775

† Conditional power threshold for the two-stage procedure.

6. Weighted Log Rank Test and Conditional Power for Adding Follow-up and Subjects

We now consider the situation in which additional subjects may be added

after the first stage of the trial. Formally, suppose that after the interim time point t_1, n' new subjects are enrolled, possibly in a staggered entry fashion, with n_1' and n_2' patients assigned to treatments 1 and 2, respectively, and with the same randomization probabilities as from stage 1, a_1 and a_2. For $k = 1, 2$, let $\bar{Y}_k'(t) = \sum_{i=n+1}^{n+n'} Y_i(t) I(Z_i = k-1)$ denote the total number of subjects at risk at calendar time $t + t_1$ among the newly added patients within each treatment arm , and $\bar{N}_k'(t) = \sum_{i=n+1}^{n+n'} N_i(t) I(Z_i = k-1)$, the total number of subjects who have failed by calendar time $t + t_1$ among the newly added patients, within each treatment arm.

The weighted log rank test, with t_2 units of additional follow-up time beyond the interim time point t_1 ($t_1 + t_2 < \tau_0$) and n' subjects added after the interim time t_1, can be expressed as

$$\begin{aligned}\tilde{L}_{t_1}(t_2) = &\int_0^{t_1+t_2} \tilde{K}(u) \frac{d\bar{N}_1(u) + I(u \le t_2) d\bar{N}_1'(u)}{\bar{Y}_1(u) + I(u \le t_2)\bar{Y}_1{}'(u)} \\ &- \int_0^{t_1+t_2} \tilde{K}(u) \frac{d\bar{N}_2(u) + I(u \le t_2) d\bar{N}_2'(u)}{\bar{Y}_2(u) + I(u \le t_2)\bar{Y}_2{}'(u)}, \end{aligned} \tag{10}$$

with normalizing predictable process

$$\begin{aligned}\tilde{K}(u) = w(u) &\left\{ \frac{n_1 + n_1' I(u \le t_2) + n_2 + n_2' I(u \le t_2)}{[n_1 + n_1' I(u \le t_2)][n_2 + n_2' I(u \le t_2)]} \right\}^{\frac{1}{2}} \\ &\times \frac{(\bar{Y}_1 + \bar{Y}_1{}' I(u \le t_2))(\bar{Y}_2 + \bar{Y}_2{}' I(u \le t_2))}{\bar{Y}_1 + \bar{Y}_1{}' I(u \le t_2) + \bar{Y}_2 + \bar{Y}_2{}' I(u \le t_2))}. \end{aligned}$$

Note that when $t_2 = 0$ or $n' = 0$, the modified log rank test (10) reduces to the original test (2).

To calculate the conditional power of this test, the asymptotic distribution of $(\tilde{L}_{t_1}(0), \tilde{L}_{t_1}(t_2)) = (L(t_1), \tilde{L}_{t_1}(t_2))$ is required. For notational simplicity, let $\tilde{\kappa}(u)$ denote the probabilistic limit of

$$\left[\frac{(n_1 + n_1' I(u \le t_2) + n_2 + n_2' I(u \le t_2))}{[n_1 + n_1' I(u \le t_2)][n_2 + n_2' I(u \le t_2)]} \right]^{\frac{1}{2}} \tilde{K}(u), \tag{11}$$

and let $\tilde{\gamma}(u)$ denote the probabilistic limit of

$$\left\{ \frac{(n_1 + n_1' I(u \le t_2))(n_2 + n_2' I(u \le t_2))}{(n_1 + n_1' I(u \le t_2) + n_2 + n_2' I(u \le t_2)} \right\}^{\frac{1}{2}} \left\{ \frac{\lambda(u|\beta)}{\lambda(u|0)} - 1 \right\}. \tag{12}$$

Further, define $\tilde{\mu}(t) = \int_0^t \tilde{\kappa}(u)\tilde{\gamma}(u) d\Lambda(u)$.

Using Martingale theory, we prove the following Proposition in the Appendix.

Proposition 2. *In addition to the assumptions of Proposition* 1*, assume further that for each* $k = 1, 2$*,*

$$\sup_{t\geq 0}\left|\frac{\bar{Y}_k'(t)}{n_k'} - \pi_k'(t)\right| \xrightarrow{p} 0 \tag{13}$$

and there exists a positive constant r *such that*

$$\frac{n_k'}{n_k} \to r. \tag{14}$$

Then as $n \to \infty$*, under the local alternatives,*

$$(L(t_1), \tilde{L}_{t_1}(t_2)) \xrightarrow{D} (L_1^\infty, L_2^\infty),$$

where $(L_1^\infty, L_2^\infty) \sim N(\mu, \Sigma)$*,* $\mu = (\mu(t_1), \tilde{\mu}(t_1 + t_2))'$*,*

$$\Sigma = \begin{pmatrix} v(t_1) & \int_0^{t_1}(1 + rI(u \leq t_2))^{-\frac{1}{2}} dv(u) \\ \int_0^{t_1}(1 + rI(u \leq t_2))^{-\frac{1}{2}} dv(u) & \tilde{v}(t_1 + t_2) \end{pmatrix},$$

and $\tilde{v}(\cdot)$ *is given in the Appendix.*

Under assumptions (13) and (14), $\tilde{\gamma}(u) = \gamma(u)$ and

$$\tilde{\kappa}(u) = \frac{w(u)\prod_{k=1}^2\{\pi_k(u) + \pi_k'(u)rI(u \leq t_2)\}}{\{1 + rI(u \leq t_2)\}\sum_{k=1}^2 a_k\{\pi_k(u) + \pi_k'(u)rI(u \leq t_2)\}}. \tag{15}$$

We note that the scientific interpretation of the weighted log rank test changes also over time. With a change in the censoring distribution over the course of the interim analyses, the weighting used to compare the hazards changes. Therefore, under the proportional hazards frailty model, the test statistics are changing functionals of the underlying distribution function (of time to event and time to censoring).

7. Designed Extension Via Conditional Power: Adding Follow-up and Subjects

Assuming further that $r = n'/n$ and recalling that $v(s) = \int_0^s [(w^2(u)\pi_1(u)\pi_2(u))/(a_1\pi_1(u) + a_2\pi_2(u))]\, du$, it follows from Proposition 2 that

$$\mathrm{Cov}(\tilde{L}_{t_1}(t_2), \tilde{L}(t_1)) \approx \begin{cases} \frac{1}{\sqrt{r+1}} v(t_1) & \text{if } t_2 > t_1 \\ \frac{1}{\sqrt{r+1}} v(t_2) - \Big[v(t_2) - v(t_1)\Big] & \text{if } 0 \leq t_2 \leq t_1. \end{cases} \tag{16}$$

Therefore, when the conditional power function is a function of both additional

follow-up, t_2 and proportion of original subjects to be added, r, it follows that

$$CP_0(t_2, r, z_\alpha | l_{t_1}) = P_0\left(\frac{\tilde{L}_{t_1}(t_2)}{\sqrt{\tilde{v}(t_1+t_2)}} > z_\alpha \middle| \frac{L(t_1)}{\sqrt{v(t_1)}} = l_{t_1}\right)$$
$$\approx 1 - \Phi\left(\frac{z_\alpha\sqrt{\tilde{v}(t_1+t_2)} - \rho\, l_{t_1}\sqrt{v(t_1)}}{\sqrt{\tilde{v}(t_1+t_2) - \rho^2 v(t_1)}}\right),$$

where $\rho = 1 - (1-(1+r)^{-1/2})\min\{1, v(t_2)/v(t_1)\}$.

When $l_{t_1} > z_\alpha$, $CP_0(t_2, r, z_\alpha | l_{t_1})$ attains the maximum value 1 at $r = 1$ and $t_2 = 0$. Otherwise $CP_0(t_2, r, z_\alpha | l_{t_1}) \doteq 1 - \Phi((\lambda z_\alpha - \rho l_{t_1})/\sqrt{\lambda^2 - \rho^2})$ where $\lambda = \sqrt{\tilde{v}(t_1+t_2)/v(t_1)}$. Note that $1 \le \lambda < [v(\infty)/v(t_1)]^{1/2}$ and $0 < \rho \le 1$. Write $f(\lambda, \rho) = (\lambda z_\alpha - \rho l_{t_1})/\sqrt{\lambda^2 - \rho^2}$. When $l_{t_1} \le 0$, it follows that $\partial f/\partial\lambda < 0$, $\partial f/\partial\rho > 0$, and $f(\lambda, \rho)$ decreases to z_α when $\lambda = 1$ (i.e., $t_2 = 0$) and $\rho \to 0$ ($r \to \infty$). When $0 < l_{t_1} < z_\alpha$, simple algebra shows that for any fixed ρ, $f(\lambda, \rho)$ is minimized at $\lambda = z_\alpha\rho/l_{t_1}$ with minimum $\sqrt{z_\alpha^2 - l_{t_1}^2}$. Similarly, for any fixed λ, $f(\lambda, \rho)$ is minimized at $\rho = l_{t_1}\lambda/z_\alpha$ with the same minimum $\sqrt{z_\alpha^2 - l_{t_1}^2}$. Therefore the minimum of $f(\lambda, \rho)$ is $\sqrt{z_\alpha^2 - l_{t_1}^2}$ and is achieved by any (λ, ρ) satisfying $\lambda/\rho = z_\alpha/l_{t_1}$. It follows that when $0 < l_{t_1} < z_\alpha$, $CP_0(t_2, r, z_\alpha | l_{t_1})$ attains the maximum $1 - \Phi(\sqrt{z_\alpha^2 - l_{t_1}^2})$ at any (t_2, r) such that $\sqrt{\tilde{v}(t_1+t_2)/v(t_1)} = (z_\alpha/l_{t_1})(1-(1-(1+r)^{-1/2})\min\{1, v(t_2)/v(t_1)\})$. Therefore, the maximum type I error for weighted log rank test that adds both follow-up and subjects is $\alpha_{\max} = \alpha + \exp(-z_\alpha^2/2)/4$, as in Proschan and Hunsberger (1995) simple scenario of testing normal means. Also, when using a larger critical value $k > z_\alpha$ and not continuing the study for $l_{t_1} < z_{p^*}$ for some $p^* \le 0.5$, the maximum type I error is $\alpha_{\max} = \int_{-\infty}^{\infty} A_{cir}(z_1)\phi(z_1)\, dz_1$, where $A_{cir}(\cdot)$ is as defined in equation (2) of Proschan and Hunsberger (1995).

This maximal error function is one example of a function, $A(l_{t_1})$. The conditional power procedure is most generally defined with respect to any such function, $A(l_{t_1})$. Now we consider the choice of t_2 and r based on any such function. Setting $CP_0(t_2, r, c | l_{t_1}) = A(l_{t_1})$ yields

$$c(t_2, r) = \frac{\rho\sqrt{v(t_1)}\, l_{t_1} + \sqrt{\tilde{v}(t_1+t_2) - \rho^2 v(t_1)}\, z_A}{\sqrt{\tilde{v}(t_1+t_2)}}. \tag{17}$$

However we choose (t_2, r), the type I error is α as long as the critical value $c(t_2, r)$ is chosen to satisfy (17). In particular, we may choose (t_2, r) such that $CP_\beta(t_2, r, c | l_{t_1}) = 1 - \delta$. Plugging (17) into the expression for $CP_\beta(t_2, r, c | l_{t_1})$, it follows that this conditional power can be achieved by any (t_2, r) such that

$$z_A + z_\delta = \frac{\tilde{\mu}(t_1+t_2) - \rho\,\mu(t_1)}{\sqrt{\tilde{v}(t_1+t_2) - \rho^2 v(t_1)}},$$

where $z_A = z_{A(l_{t_1})}$. Note that this reduces to (9) when new subjects are not added (i.e., when $r = 0$ and therefore $\rho = 1$).

8. Application to Frailty Model

Although motivated by the problem of heterogeneity in randomized trials, many components of the results presented in Sections $3 - 7$ are applicable to any situation in which the weighted log rank test is used for analysis. For application of the conditional power procedure, a model must be assumed for the data and the unknown parameters of the model must be known. The analysis at t_1 can typically be used to estimate these parameters. For the frailty model, these include the baseline hazard function and the frailty variance parameter, θ. Although θ is not identifiable in the absence of additional data (at least in the presence of a binary treatment indicator), in many situations it may be reasonable to assume that there is a secondary endpoint that shares the frailty with the endpoint of interest (Oakes (1989)). In this case, θ is estimable from the marginal association of the two events. For example, for a log gamma frailty, $\theta = 1/2\tau - 1$, where τ is Kendall's tau measure of association between the two events. Thus, a simple nonparametric estimate of θ is available at the stage 1 analysis that does not depend on the joint survivor distribution of the two events. If one of the endpoints is death, and thus the event times are ordered, the work of Jiang, Fine and Chapell (1999) is applicable. Of potential concern is the variability with which the measure of association is estimated, especially at a relatively early point in the study. However θ is estimated, it must be done so consistently so that Slutsky's Theorem can be invoked. We recommend a sensitivity analysis with respect to θ, informed by the estimated measure of association.

9. Brain Tumor Example

To illustrate the proposed design, we use an expanded version of a brain tumor data set analyzed in Ino et al. (2001). Patients in this study were diagnosed with anaplastic oligodendroglioma, a common variant of malignant brain tumor. Ninety-six subjects were enrolled in the study, and 49 of them died during the study period. Loss of Heterozygosity (LOH) of chromosome 1p was found to be highly predictive of survival among these patients. For illustration, we consider the problem of testing whether LOH at chromosome 19q is predictive of survival when ignoring LOH at 1p; that is, we pretend that the biologists had not yet discovered that 1p LOH is predictive of survival and thus artificially ensure that there is important unrecognized heterogeneity in the data. Indeed in this study a number of subjects were missing the covariate of 1p LOH.

We consider what would have happened if the study had been conducted in two stages, the first being at $t_1 = 185$ months. For simplicity, we assume that the

frailty, b, induced by ignoring LOH at 1p, has a known variance $\hat{\theta} = 0.571$. We used $\hat{p}(1-\hat{p})\hat{\eta}^2$ to estimate this variance, where $\hat{p}$ is the proportion of subjects with 1p LOH, which is estimated based on subjects with complete 1p LOH information, and $\hat{\eta}$ is the estimated coefficient for 1p LOH in a proportional hazards model with covariates 1p LOH and 19q LOH among subjects with complete 1p LOH and 19q LOH information. To account for the variability of the estimated frailty parameter, we propose a sensitivity analysis to examine the choice of $\hat{\theta}$. This sensitivity analysis approach is immediately available to the analysis of real data for which we suspect heterogeneity, but the magnitude of such heterogeneity is uncertain.

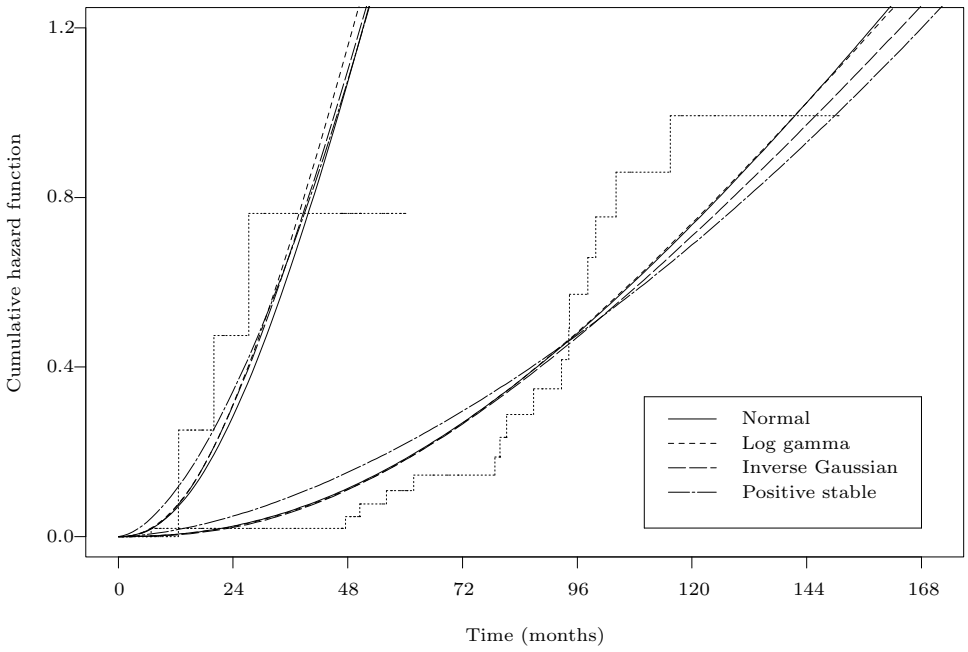

Figure 1. Estimated cumulative hazard functions under various frailty distributions. The step functions in dotted lines represent nonparametric estimates of the cumulative hazard functions.

At $t_1 = 185$ months, 66 subjects had entered the study: 54 of them (with 16 deaths) had 19q LOH and 12 of them (with 4 deaths) did not have 19q LOH. The fitted cumulative hazard functions with a Weibull baseline hazard are shown in Figure 1 for the four parametric frailty models considered in Section 2. Also shown in Figure 1 are the nonparametric estimates of the cumulative hazard functions (dotted lines). All frailty models provide a reasonably good fit, with

the fit under the positive stable model worse than that of the other three. The positive stable model yields proportional marginal hazard functions and thus does not capture the nonproportional marginal hazards observed in this dataset.

For illustration, we carried out the conditional power calculation for the weighted log rank test. Under the normal frailty model, the optimally weighted log rank statistic is $L(t_1)/\sqrt{v(t_1)} = -1.516$, with one-sided p-value 0.065. The optimal weight is defined to maximize the Pitman asymptotic efficacy defined in (4.11) of Fleming and Harrington (1991, p.270)). The variance function, $v(t_1)$, is calculated under the assumed parametric model

$$\lambda(t|LOH19q, b) = \lambda_0 t^{p-1} \exp(\beta \cdot 1_{\{LOH19q=1\}} + b),$$

where the frailty b follows a normal distribution with mean 0 and variance 0.571. The maximum likelihood estimates for λ_0, p and β are 4.23×10^{-4}, 2.244, and -2.525, respectively. It follows that $R_{\max} = 0.441$. For $p^* = 0.15$, we have $k = 1.770$ for $r = 0$ and the conditional error function derived in Section 4, and we have $k = 1.821$ for $r > 0$ and the circular conditional error function, A_{cir} (Section 7). Figure 2(a) displays the conditional power for extending follow-up time by 6 to 36 months and adding r times as many new subjects, assuming that 80% of the new subjects enter the study uniformly during the first one-third of the follow-up period and the other 20% of new subjects enter the study uniformly during the last two-thirds of the follow-up period. It is clear that extending follow-up alone will not yield high conditional power, as can be expected from moderate $R_{\max}$. To attain 80% conditional power, one can either add half as many subjects ($r = 0.5$) and extend the follow-up by 27.4 months, add as many subjects ($r = 1$) and extend the follow-up by 22.6 months, or add twice as many subjects ($r = 2$) and extend the follow-up by 18.6 months. Also shown in Figure 2 are the conditional powers of the weighted log rank test under the log gamma, inverse Gaussian, and positive stable frailty models. The conditional powers for the log gamma and inverse Gaussian frailty model are similar to that for a normal frailty model. For example, to obtain 80% conditional power with as many new subjects, it requires extending follow-up by 22.9 months under a log gamma frailty model and 24.7 months under an inverse Gaussian frailty model, as compared to 22.9 months under a normal frailty model. For the not as well fitted positive stable frailty model, it requires 27.0 months of additional follow-up. These results suggest that the designed extension using conditional power is robust to the selected frailty model, as long as the chosen frailty model provides a reasonable fit to data.

We went through this same exercise for the simple log rank statistic. With adding as many new subjects ($r = 1$), the required follow-up for the normal, log

gamma, inverse Gaussian and positive stable frailty models are 29.6, 34.4, 34.2, and 27.0 months, respectively; all of them are longer than those for the weighted log rank statistic. This suggests that in spite of the added computational burden of computing the optimal weights, the weighted test is preferable because of the efficiency gained.

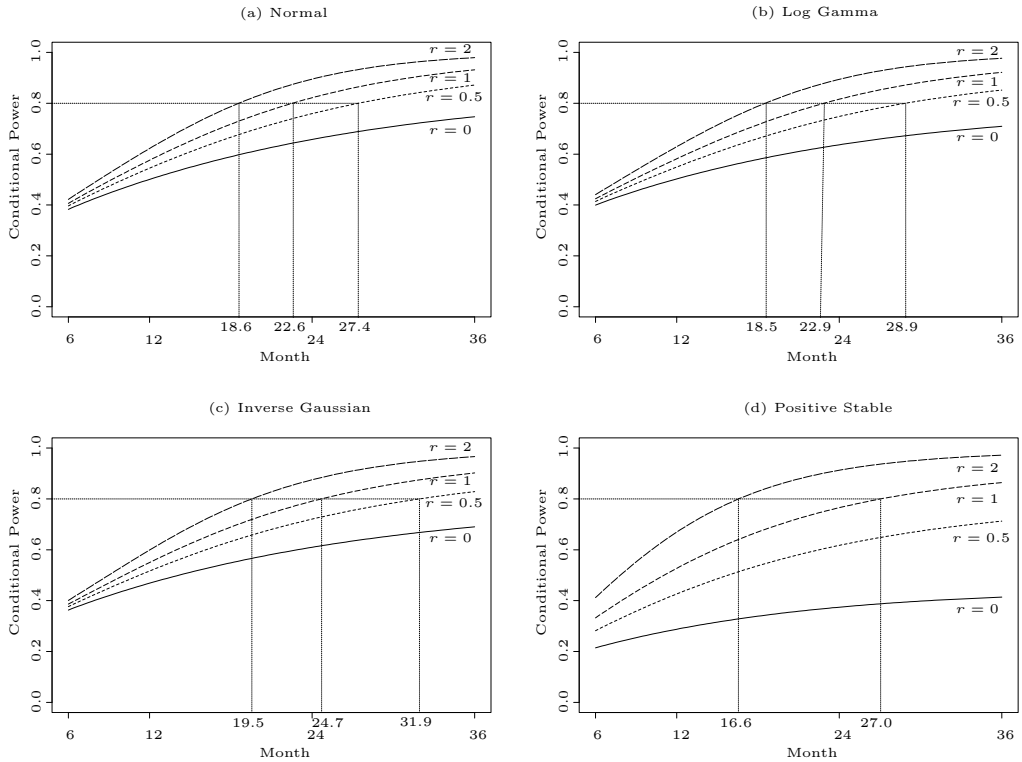

Figure 2. Conditional power based on the optimally weighted log rank statistic for extending follow-up time (in months) and adding r times as many new subjects, under various frailty distributions: (a) normal; (b) log gamma; (c) inverse Gaussian; (d) positive stable.

Figure 3 gives the conditional power under the log gamma frailty model when adding one times as many new subjects in stage 2. The conditional power is similar for θ ranging from 0.4 to 0.8. In particular, compared to 22.9 months for $\theta = 0.571$, the follow-up time required to achieve 80% conditional power for $\theta = 0.4, 0.5, 0.7, 0.8$ are 26.2, 24.1, 21.0 and 20.0 months, respectively. Based on this sensitivity analysis, one might choose the more conservative estimate of 26 months.

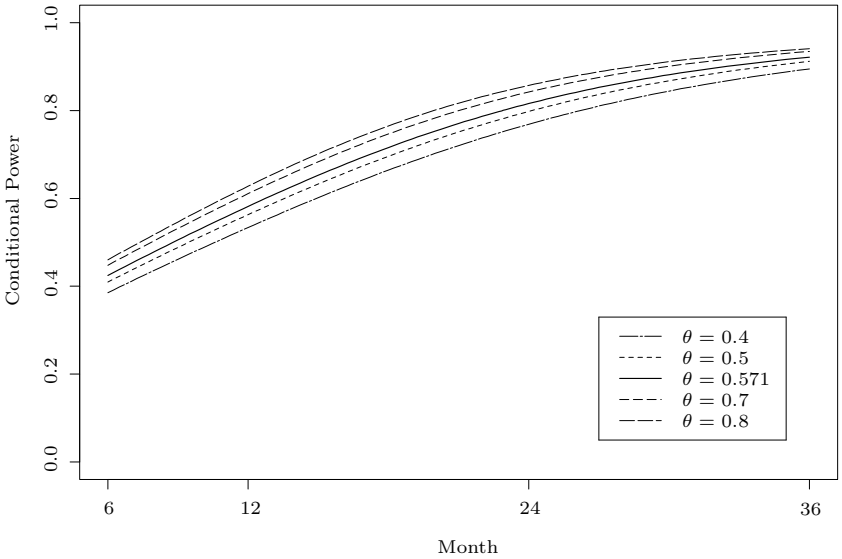

Figure 3. Conditional power based on the optimally weighted log rank statistic for extending follow-up time (in months) and adding one time as many new subjects, under the log gamma frailty distribution and a range of θ values.

10. Discussion

We have proposed an adaptive procedure for a survival endpoint to recover power when the model assumed for design of the trial is suspect. It is worth noting that, as opposed to Proschan and Hunsberger (1995) who confined the application of conditional power design to parametric tests, we have generalized the applications to semi-parametric tests for data which are subject to censoring. Our motivating application was the common situation of unrecognized genetic heterogeneity, known to dilute the power of the log rank test through decreased precision. This procedure allows for extension of follow-up time, with or without additional accrual of subjects. It also allows for early stopping for futility if the treatment is ineffective. In this regard, it is widely applicable as it may be tailored to the relative costs of the study.

Effect modification is another plausible mechanism for loss of power. This occurs if the effect of chemotherapy varies by genetic subtype. This would require including an interaction term in the proportional hazards frailty model. This may lead to a non-monotone drift function, making the conditional power procedure inapplicable (i.e., equation (9) cannot be solved). Thus, alternative adaptive designs are required for this situation.

As for any adaptive procedure, estimates of unknown parameters at early points in the study may be highly variable. This is especially true for estimation of θ as a shared frailty between event times, and for the baseline hazard function. A partial remedy to this problem would be to re-estimate the frailty parameter at the end of the study and to use it in the final analysis. The impact of this approach on the overall type I error remains an interesting open question.

It must be recognized that this procedure addresses the problem of unexplained heterogeneity at the cost of modeling assumptions: the frailty proportional hazards model, including the baseline hazard function and the frailty distribution, and the shared frailty model. All of these assumptions certainly call for sensitivity analyses and conservative choices. If genetic (or other) information becomes available during the course of the trial, it would be desirable to use that information in the selection of subjects. However, short of new discoveries during the course of the study that explain heterogeneity, an adaptive procedure such as this to recover power may be the best approach.

Acknowledgements

This research was supported in part by NIH grants CA75971, CA57253, CA95747 and AI24643. We thank D. Louis and J.G. Cairncross for the use of the brain tumor data. We thank the Editor, an associate editor and two referees for insightful suggestions, which significantly improved the presentation of this manuscript.

Appendix: Proof of Proposition 2.

We prove the weak convergence theorem by using the Cramer-Wold device; we show that any linear combination of $L(t_1)$ and $\tilde{L}_{t_1}(t_2)$, under the local alternatives, converges weakly to the same linear combination of L_1^∞ and L_2^∞, as $n \to \infty$.

We may re-express $L(t_1)$ and $\tilde{L}_{t_1}(t_2)$ as

$$\begin{aligned} L(t_1) &= \int_0^{t_1} K(u)\frac{dM_1}{\bar{Y}_1} - \int_0^{t_1} K(u)\frac{dM_2}{\bar{Y}_2} + \int_0^{t_1} K(u)(d\Lambda_{2,n} - d\Lambda_{1,n}) \\ &= \int_0^{t_1+t_2} K(u)I(u \le t_1)\frac{dM_1}{\bar{Y}_1} - \int_0^{t_1+t_2} K(u)I(u \le t_1)\frac{dM_2}{\bar{Y}_2} \\ &\quad + \int_0^{t_1} K(u)(d\Lambda_{2,n} - d\Lambda_{1,n}), \end{aligned}$$

$$\begin{aligned} \tilde{L}_{t_1}(t_2) &= \int_0^{t_1+t_2} \tilde{K}(u)\frac{dM_1 + I(u \le t_2)dM_1'}{\bar{Y}_1 + I(u \le t_2)\bar{Y}_1'} - \int_0^{t_1+t_2} \tilde{K}(u)\frac{dM_2 + I(u \le t_2)dM_2'}{\bar{Y}_2 + I(u \le t_2)\bar{Y}_2'} \\ &\quad + \int_0^{t_1+t_2} \tilde{K}(u)(d\Lambda_{2,n} - d\Lambda_{1,n}), \end{aligned}$$

where, for $k = 1, 2$, $M_k(u) = N_k(u) - \int_0^u \bar{Y}_k(s)d\Lambda_{k,n}(s)$ and $M_k'(u) = N_k'(u) - \int_0^u \bar{Y}_k'(s)d\Lambda_{k,n}(s)$ are martingales with respect to the filtration $\mathcal{F}(u) = \sigma(N_k(s), \bar{Y}_k(s), N_k'(s), \bar{Y}_k'(s), 0 \le s \le u, k = 1, 2)$, and $\Lambda_{k,n}$ is the marginal cumulative hazard for treatment arm k. Note that $\Lambda_{1,n}(u) \equiv \Lambda(u)$.

Hence, for any two real numbers, c_1, c_2,

$$
\begin{aligned}
&c_1 L(t_1) + c_2 \tilde{L}_{t_1}(t_2) \\
&\quad = \int_0^{t_1+t_2} \left\{ \frac{c_1 K(u) I(u \le t_1)}{\bar{Y}_1} + \frac{c_2 \tilde{K}(u)}{\bar{Y}_1 + I(u \le t_2)\bar{Y}_1'} \right\} dM_1(u) \\
&\qquad + \int_0^{t_1+t_2} \frac{c_2 \tilde{K}(u) I(u \le t_2) dM_1'}{\bar{Y}_1 + I(u \le t_2)\bar{Y}_1'} \\
&\qquad - \int_0^{t_1+t_2} \left\{ \frac{c_1 K(u) I(u \le t_1)}{\bar{Y}_2} + \frac{c_2 \tilde{K}(u)}{\bar{Y}_2 + I(u \le t_2)\bar{Y}_2'} \right\} dM_2(u) \\
&\qquad - \int_0^{t_1+t_2} \frac{c_2 \tilde{K}(u) I(u \le t_2) dM_2'}{\bar{Y}_2 + I(u \le t_2)\bar{Y}_2'} \\
&\qquad + c_1 \int_0^{t_1} K(u)(d\Lambda_{2,n} - d\Lambda_{1,n}) + c_2 \int_0^{t_1+t_2} \tilde{K}(u)(d\Lambda_{2,n} - d\Lambda_{1,n}). \qquad (18)
\end{aligned}
$$

The last two terms in (18) converge in probability to

$$
c_1 \int_0^{t_1} \kappa(u)\gamma(u)d\Lambda(u) + c_2 \int_0^{t_1+t_2} \tilde{\kappa}(u)\tilde{\gamma}(u)d\Lambda(u),
$$

where $\tilde{\kappa}(u)$ and $\tilde{\gamma}(u)$ are as defined in (11), (12) and (15).

Noticing that M_1, M_1', M_2, M_2' are independent martingales, we consider square integrable martingale terms

$$
\begin{aligned}
&\int_0^{t_1+t_2} H_1(u)dM_1(u), \int_0^{t_1+t_2} H_1'(u)dM_1'(u), \\
&\int_0^{t_1+t_2} H_2(u)dM_2(u), \int_0^{t_1+t_2} H_2'(u)dM_2(u)
\end{aligned}
$$

in (18), where for $k = 1, 2$,

$$
\begin{aligned}
H_k(u) &= \frac{c_1 K(u) I(u \le t_1)}{\bar{Y}_k} + \frac{c_2 \tilde{K}(u)}{\bar{Y}_k + I(u \le t_2)\bar{Y}_k'}, \\
H_k'(u) &= \frac{c_2 \tilde{K}(u) I(u \le t_2)}{\bar{Y}_1 + I(u \le t_2)\bar{Y}_1'}.
\end{aligned}
$$

With assumptions (3) − (14), it follows that $\sup_{0<u\le t_1+t_2} |\{H_k(u)\}^2 \bar{Y}_k(u) -$

$h_k(u)| \xrightarrow{p} 0$ and $\sup_{0<u\le t_1+t_2} |\{H_k'(u)\}^2 \bar{Y}_k'(u) - h_k'(u)| \xrightarrow{p} 0$, where

$$h_k(u) = \left(c_1 I(u \le t_1) a_{3-k}^{\frac{1}{2}} \pi_k^{-\frac{1}{2}} \kappa(u) + c_2 \{1 + r I(u \le t_2)\}^{\frac{1}{2}} \frac{a_{3-k}^{\frac{1}{2}} \pi_k^{\frac{1}{2}}}{\pi_k + r\pi_k' I(u \le t_2)} \tilde{\kappa}(u) \right)^2,$$

$$h_k'(u) = I(u \le t_2) c_2^2 \frac{r(1+r) a_{3-k} \pi_k}{\{\pi_k + r\pi_k' I(u \le t_2)\}^2} \tilde{\kappa}^2(u).$$

By applying the Martingale Central Limit Theorem (e.g., Theorem 6.2.1 in Fleming and Harrington, (1991)), it follows that

$$\left(\int_0^{t_1+t_2} H_1(u) dM_1(u), \int_0^{t_1+t_2} H_1'(u) dM_1'(u), \int_0^{t_1+t_2} H_2(u) dM_2(u), \right.$$
$$\left. \int_0^{t_1+t_2} H_2'(u) dM_2(u) \right) \xrightarrow{D} (\mathcal{L}_1, \mathcal{L}_1', \mathcal{L}_2, \mathcal{L}_2'),$$

where $\mathcal{L}_1, \mathcal{L}_1', \mathcal{L}_2, \mathcal{L}_2'$ are mutually independent mean zero normal random variables with variances $\int_0^{t_1+t_2} h_1(u)\, d\Lambda(u)$, $\int_0^{t_1+t_2} h_1'(u)\, d\Lambda(u)$, $\int_0^{t_1+t_2} h_2(u)\, d\Lambda(u)$, $\int_0^{t_1+t_2} h_2'(u)\, d\Lambda(u)$, respectively. Some algebra then shows that the first four terms in (18) converge weakly to a Gaussian random variable with mean 0 and variance

$$c_1^2 v(t_1) + c_2^2 \tilde{v}(t_1+t_2) + 2c_1c_2 \int_0^{t_1} \{1 + r I(u \le t_2)\}^{-1/2} dv(u),$$

where $\tilde{v}(t_1+t_2) = \int_0^{t_1+t_2} w(u)\tilde{\kappa}(u) d\Lambda(u)$.

Hence, by Slutsky's theorem, $c_1 L(t_1) + c_2 \tilde{L}_{t_1}(t_2) \xrightarrow{D} c_1 L_1^\infty + c_2 L_2^\infty$.

References

Betensky, R. A. (1998). Construction of a continuous stopping boundary from an alpha spending function. *Biometrics* **54**, 1061-1071.

Betensky, R. A., Louis D. N. and Cairncross, J. G. (2002). The influence of unrecognized molecular heterogeneity on randomized clinical trials. *J. Clinical Oncology* **20**, 2495-2499.

Betensky, R. A., Cairncross, J. G. and Louis, D. N. (2003). Analysis of a molecular genetic neuro-oncology study with partially biased selection. *Biostatistics* **4**, 167-178.

Fleming, T. R. and Harrington D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

Hougaard, P. (1984). Life table methods for heterogeneous populations: distributions describing the heterogeneity. *Biometrika* **71**, 75-83.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from positive stable distributions. *Biometrika* **73**, 387-396.

Ino, Y., Betensky, R. A., Zlatescu, M. C., Sasaki, H., Macdonald, D. R., Stemmer-Rachamimov, A. O., Ramsay, D. A., Cairncross, J. G. and Louis, D. N. (2001). Molecular subtypes of anaplastic oligodendroglioma: implications for patient management at diagnosis. *Clinical Cancer Research* **7**, 839-845.

Jiang, H., Fine, J. and Chapell, R. (1999). Estimation of the association between bivariate failure times in semi-competing risk problems. *ASA Proceedings of the Biometrics Section*, 192-194.

Li, Y., Betensky, R. A., Louis, D. N. and Cairncross, J. G. (2002). The use of frailty hazard models for unrecognized heterogeneity that interacts with treatment: considerations of efficiency and power. *Biometrics* **58**, 232-236.

Oakes, D. (1989). Bivariate survival models induced by frailties. *J. Amer. Statist. Assoc.* **84**, 487-493.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315-1324.

Schaefer, H. and Mueller, H. H. (2001). Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statist. Medicine* **20**, 3741-3751.

Shen, Y. and Cai, J. W. (2003). Sample size reestimation for clinical trials with censored survival data. *J. Amer. Statist. Assoc.* **98**, 418-426.

Dana-Farber Cancer Institute, Harvard University, 44 Binney Street, LW211, Boston, MA 02115, U.S.A.

E-mail: yili@jimmy.harvard.edu

Department of Health Research and Policy, HRP Redwood Building, Standford University, School of Medicine, Stanford, CA 94305-5405, U.S.A.

E-mail: meichiun@standford.edu

Harvard School of Public Health, Harvard University, 655 Huntington Avenne, Boston, MA 02115, U.S.A.

E-mail: betensky@hsph.harvard.edu

(Received February 2005; accepted February 2006)