BIOMETRICS 58, 232–236
March 2002

The Use of Frailty Hazard Models for Unrecognized Heterogeneity That Interacts with Treatment: Considerations of Efficiency and Power

Yi Li,[1,2] Rebecca A. Betensky,[1,*] David N. Louis,[3] and J. Gregory Cairncross[4]

[1] Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A.
[2]Department of Biostatistics, Dana-Farber Cancer Institute, Boston, Massachusetts, U.S.A.
[3]Department of Pathology and Neurosurgical Service, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, U.S.A.
[4]Department of Oncology, University of Western Ontario and London Regional Cancer Centre, London, Ontario, Canada
email: betensky@hsph.harvard.edu

SUMMARY. Increasingly, genetic studies of tumors of the same histologic diagnosis are elucidating subtypes that are distinct with respect to clinical endpoints such as response to treatment and survival. This raises concerns about the efficiency of using the simple log-rank test for analysis of treatment effect on survival in studies of possibly heterogeneous tumors. Furthermore, such studies, designed under the assumption of homogeneity, may be severely underpowered. We derive analytic approximations for the asymptotic relative efficiency of the simple log-rank test relative to the optimally weighted log-rank test and for the power of the simple log-rank test when applied to subjects with unobserved heterogeneity, as reflected in a continuous frailty, that may interact with treatment. Numerical studies demonstrate that the simple log-rank test may be quite inefficient if the frailty interacts with treatment. Further, there may be a substantial loss of power in the presence of the frailty with or without an interaction with treatment.

KEY WORD: Omitted covariate.

1. Introduction

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, are currently nearing completion. These trials were designed prior to the discovery of at least three clinically distinct genetic subtypes among patients with the histological diagnosis of this disease (Ino et al., 2001). Patients with allelic loss of chromosome 1p respond to chemotherapy and have long survival times. In contrast, patients with chromosome 1p intact and no mutation of the TP53 gene respond infrequently to chemotherapy and have short survival times. Patients with chromosome 1p intact and a TP53 mutation follow an intermediate course. 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.

Within the context of the proportional hazards model, several authors have considered the problem of unrecognized heterogeneity that divides the patients into two distinct groups (e.g., Lagakos and Schoenfeld, 1984; Struthers and Kalbfleisch, 1986; Chastang, Byar, and Piantadosi, 1988; Schmoor and Schumacher, 1997). In statistical terms, this amounts to an omitted binary covariate. Almost all of these articles have assumed that the omitted covariate is independent of the treatment assignment and have precluded the possibility of an interaction between treatment and the omitted covariate. An exception to this is presented by Lagakos and Schoenfeld (1984), in which such an interaction is allowed. As in the example motivating this work, there will frequently be an interaction between treatment and the omitted covariate when it describes a genetic subtype.

Another important extension of the omitted binary covariate problem is to assume that heterogeneity confers a continuum of relative risk. This is more plausible than the assumption of two underlying subgroups within the disease diagnosis, as there are likely to be several as yet unrecognized molecular or other features that operate jointly to confer different risks. A natural way to allow for this is to assume that each individual has a distinct realization of a frailty that modifies his or her hazard for death. This approach has been considered for the case in which the frailty does not interact with

treatment by Keiding and Andersen (1997) for a Weibull accelerated failure time model and by Lagakos and Schoenfeld (1984) for the proportional hazards model. In this article, we extend the results of Lagakos and Schoenfeld (1984) to examine the impact of unmeasured heterogeneity, as captured through a continuous frailty that interacts with treatment, on the efficiency and power of the simple log-rank test for treatment effect.

There are two distinct concepts of asymptotic relative efficiency in the literature on omitted covariates, which are often used interchangeably, with ensuing confusion. The more common concept relates the simple log-rank test to the optimally weighted log-rank test assuming the model of unobserved heterogeneity to be true. This concept was used, e.g., by Lagakos and Schoenfeld (1984, Appendix 1) and Oakes and Jeong (1998). This measure is useful, as it is informative about the loss of efficiency in using the simple log-rank test relative to the more complicated optimally weighted log-rank test in a situation in which there is heterogeneity. The second concept of efficiency compares the simple log-rank test when applied to data without heterogeneity with the same test when applied to data with unobserved heterogeneity. It was used by Lagakos and Schoenfeld (1984, Appendix 3) and Morgan (1986). This concept is useful when evaluating the power of the simple log-rank test for a study that was designed assuming there to be no heterogeneity, when in fact there is heterogeneity (as in Betensky et al., unpublished manuscript). We evaluate both of these types of efficiency for the scenario that we consider because they each provide useful and complementary information. To avoid confusion, we translate the second measure from that of efficiency into the actual power of the simple log-rank test. All future references to efficiency imply the first usage.

In Section 2, we introduce notation and the models. In Section 3, we relate the parameters of the naive model that ignores the frailty to those of the correct model that incorporates the frailty. In Section 4, we derive an analytic expression for the asymptotic relative efficiency (ARE) of the simple log-rank test under the true model of heterogeneity. In Section 5, we derive an analytic approximation for the power of the simple log-rank test when applied to heterogeneous data. We conclude in Section 6 with numerical illustrations of these results.

2. Notation and Models

Suppose each of n patients is randomly assigned to one of the two treatment arms according to the distribution P, leading to n_1 patients receiving treatment 1 and n_2 patients receiving treatment 2. Let Z indicate the treatment received, with $Z = 0$ for patients who receive treatment 1 and $Z = 1$ for patients who receive treatment 2. Associated with each subject is a death time, T_i, and a censoring time, C_i, distributed according to G such that $G(t) = P(C > t)$. However, only $X_i = \min(T_i, C_i)$ is observed, along with an indicator for whether death was observed, $\delta_i = 1$ if $T_i \leq C_i$ and $\delta_i = 0$ if $T_i > C_i$. Associated with each subject is an unobserved frailty, b, where $b \sim F(b;\theta)$ and θ parameterizes the frailty distribution, F.

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

$$\lambda(t \mid Z, b) = \lambda_0(t)\exp(\beta Z + b + \alpha b Z). \tag{1}$$

Here β measures the main effect for treatment and α measures the effect of the interaction between the frailty and treatment. If the heterogeneity among the subjects is not recognized, the naive proportional hazards model,

$$\lambda^*(t \mid Z) = \lambda_0^*(t)\exp(\beta^* Z), \tag{2}$$

is assumed instead.

3. Heuristic Connection Between the Models

Following derivations in the measurement error literature (e.g., Carroll et al., 1995), the induced hazard function based on the true model (1) is given by

$$\lambda(t \mid Z) = \lambda_0(t)\mathrm{E}[\exp(\beta Z + b + \alpha b Z) \mid T \geq t,\, Z].$$

The expectation in this expression involves the baseline hazard function, $\lambda_0(t)$, which complicates its evaluation. However, if the event is rare and $\{T \geq t\}$ occurs with high probability, and assuming that $b \sim \mathrm{N}(0, \sigma^2)$, the induced hazard function can be approximated as

$$\begin{aligned}\lambda(t \mid Z) &= \lambda_0(t)\mathrm{E}[\exp(\beta Z + b + \alpha b Z) \mid Z] \\ &= \lambda_0(t)\exp(\sigma^2/2)\exp\left[\left(\beta + \sigma^2\alpha + \sigma^2\alpha^2/2\right) Z\right].\end{aligned}$$

Note that the induced hazard function is now of exactly the same form as the naive hazard function (2), with β^* corresponding to $\beta + \sigma^2\alpha + \sigma^2\alpha^2/2$. Thus, the treatment effect based on the naive model is a function of the treatment effect from the correct model, (β, α), as well as the variance of the frailty.

This correspondence illustrates that rejection of the naive null hypothesis for no treatment effect, H_0: $\beta^* = 0$, implies rejection of the correct null hypothesis for no treatment effect, H_0: $\beta = 0, \alpha = 0$. However, it is possible for neither β nor α to be zero and yet for β^* to be zero, i.e., if $\alpha = -0.5 \pm 0.5(1 - 2\beta/\sigma^2)^{1/2}$. Thus, it is apparent, heuristically, through some simplifying assumptions that the simple log-rank test based on the naive model (2) will be inefficient relative to the optimally weighted log-rank test. We explore this more formally in the next section.

4. ARE of the Simple Log-Rank Test

Here we derive the asymptotic relative efficiency (ARE) of the simple log-rank test (i.e., the score test based on model (2)) versus the optimally weighted log-rank test for a treatment effect based on data that follow the proportional hazards frailty model with a treatment × frailty interaction (1). The log-rank test can be written in the following general form:

$$W = \int K\frac{d\bar{N}_1}{\bar{Y}_1} - \int K\frac{d\bar{N}_2}{\bar{Y}_2}, \tag{3}$$

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

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

In the simple log-rank test, $w(s) \equiv 1$.

Under model (1), the marginal hazard function for the $Z = 1$ group is

$$\lambda(t \mid \beta, \alpha) = \lambda_0(t)\exp(\beta)\frac{-P'_\alpha\{\Lambda_\beta(t)\}}{P_\alpha\{\Lambda_\beta(t)\}},$$

where $\Lambda_\beta(t) = \Lambda_0(t)\exp(\beta)$ and $P_\alpha(s) = \int e^{-s\exp(b+\alpha b)}dF(b)$ is the Laplace transform for the random variable $\exp(b+\alpha b)$. Note that $\lambda(t;0,0)$ is the marginal hazard function for the $Z = 0$ group. Further, assuming differentiability of $\lambda(t \mid \beta, \alpha)$ with respect to β, α in a neighborhood of $(0,0)$, a Taylor series expansion yields

$$\lambda(t \mid \beta, \alpha) \doteq \lambda(t \mid 0,0) + \left.\frac{\partial}{\partial\beta}\right|_{\substack{\beta=0\\ \alpha=0}} \lambda(t \mid \beta, \alpha)\beta + \left.\frac{\partial}{\partial\alpha}\right|_{\substack{\beta=0\\ \alpha=0}} \lambda(t \mid \beta, \alpha)\alpha.$$

To examine the behavior of the ARE, we consider a sequence of alternatives that converges to the null hypothesis at the appropriate rate as sample size increases to infinity. Under such alternatives, the log-rank statistic has asymptotically a finite mean and variance. Assuming some regularity conditions and such a sequence of local alternatives, the weighted log-rank test (3) converges to

$$\mathrm{N}\left(\int_0^\infty \kappa\gamma d\Lambda, \int_0^\infty \frac{a_1\pi_1 + a_2\pi_2}{\pi_1\pi_2}\kappa^2 d\Lambda\right),$$

where κ is the probabilistic limit of $\{(n_1+n_2)/(n_1n_2)\}^{1/2}K(t)$, γ is the probabilistic limit of $\{(n_1n_2)/(n_1+n_2)\}^{1/2}[(\lambda(t \mid \beta,\alpha)/\lambda(t;0,0)) - 1]$ for $k = 1,2$, $a_k = P(Z_i = k-1)$, π_k is the probabilistic limit of $\bar{Y}_k/n_k$, and $\Lambda(t) = \int_0^t \lambda(s \mid Z,0,0)ds$. Hence, its asymptotic efficacy is defined by its noncentrality, given by

$$\frac{\left(\int_0^\infty \kappa\gamma d\Lambda\right)^2}{\int \frac{a_1\pi_1 + a_2\pi_2}{\pi_1\pi_2}\kappa^2 d\Lambda}.$$

If we specify the sequence of alternatives as

$$\mathrm{H}_a\colon \beta = \alpha = \left(\frac{n_1+n_2}{n_1n_2}\right)^{1/2},$$

it then follows that

$$\gamma(t) = \lim_{n\to\infty}\left(\frac{n_1n_2}{n_1+n_2}\right)^{1/2}\left(\frac{\lambda(t \mid \beta,\alpha)}{\lambda(t \mid 0,0)} - 1\right) = \left.\frac{\partial}{\partial\beta}\right|_{\substack{\beta=0\\ \alpha=0}} \log\lambda(t \mid \beta,\alpha) + \left.\frac{\partial}{\partial\alpha}\right|_{\substack{\beta=0\\ \alpha=0}} \log\lambda(t \mid \beta,\alpha).$$

Hence, by the Cauchy–Schwarz inequality, the optimal efficiency is achieved by taking $w(t) \propto \gamma(t)$, in which case the efficacy is

$$\int \frac{\pi_1\pi_2}{a_1\pi_1 + a_2\pi_2}\gamma^2(t)d\Lambda(t).$$

Therefore, the ARE comparing the log-rank test to the optimally weighted test is

$$\frac{\left(\int \frac{\pi_1\pi_2}{a_1\pi_1 + a_2\pi_2}\gamma(t)d\Lambda(t)\right)^2}{\int \frac{\pi_1\pi_2}{a_1\pi_1 + a_2\pi_2}d\Lambda(t)\int \frac{\pi_1\pi_2}{a_1\pi_1 + a_2\pi_2}\gamma^2(t)d\Lambda(t)}.$$

Assuming that the censoring distribution is the same for both treatment groups, i.e., $P(C > t \mid Z) = P(C > t) = G(t)$, it follows that $\pi_1(t) = \pi_2(t) = \pi(t) = G(t)\exp(-\Lambda(t))$. This leads to a simpler form of ARE given by

$$\frac{\left(\int \pi(t)\gamma(t)d\Lambda(t)\right)^2}{\int \pi(t)d\Lambda(t)\int \pi(t)\gamma^2(t)d\Lambda(t)}.$$

5. Power of the Simple Log-Rank Test

For a sequence of local alternatives (i.e., $\beta = O(n^{-1/2})$), the distribution of the simple log-rank test under the naive model (2) is approximately given by

$$\mathrm{N}\left(\left[\frac{n_1n_2}{n_1+n_2}\right]^{1/2}\int \frac{\pi_1^*(t)\pi_2^*(t)}{a_1\pi_1^*(t) + a_2\pi_2^*(t)}\beta^* d\Lambda_0^*(t), \int \frac{\pi_1^*(t)\pi_2^*(t)}{a_1\pi_1^*(t) + a_2\pi_2^*(t)}d\Lambda_0^*(t)\right),$$

where $\Lambda_0^*(t) = \int \lambda^*(s)ds$ and π_k^* is the probabilistic limit of $\bar{Y}/n_k$ under model (2). Similarly, the approximate distribution of the simple log-rank test under the true model (1) is given by

$$\mathrm{N}\left(\left[\frac{n_1n_2}{n_1+n_2}\right]^{1/2}\int \frac{\pi_1(t)\pi_2(t)}{a_1\pi_1(t) + a_2\pi_2(t)}V(t,\beta,\alpha)d\Lambda(t), \int \frac{\pi_1(t)\pi_2(t)}{a_1\pi_1(t) + a_2\pi_2(t)}d\Lambda(t)\right),$$

where

$$V(t,\beta,\alpha) = \beta \times \left.\frac{\partial}{\partial\beta}\right|_{\substack{\beta=0\\ \alpha=0}} \log\lambda(t \mid \beta,\alpha) + \alpha \times \left.\frac{\partial}{\partial\alpha}\right|_{\substack{\beta=0\\ \alpha=0}} \log\lambda(t \mid \beta,\alpha).$$

We assume for simplicity that the two treatment groups have the same censoring distributions and that there is equal allocation of treatments (i.e., $a_1 = a_2 = 0.5$). It then follows that the sample size needed to detect $\mathrm{H}_a\colon \beta^* = \beta > 0$ versus $\mathrm{H}_0\colon \beta^* = 0$ under the naive model (2) with power δ and one-sided type I error level of ϵ is

$$N_{\text{naive}} = \frac{4(Z_{1-\epsilon} + Z_\delta)^2}{\beta^2\int \pi^*(t)d\Lambda_0^*(t)},$$

where Z_q is the $100 \times q$ percentile of a standard normal distribution. Thus, for a study designed under the incorrect assumption of the naive model (2) when in fact the model of heterogeneity (1) holds, the actual power is

$$1 - \Phi\left(Z_{1-\epsilon} - 0.5N_{\text{naive}}^{1/2}\frac{\int \pi(t)V(t,\beta,\alpha)d\Lambda(t)}{(\int \pi(t)d\Lambda(t))^{1/2}}\right),$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

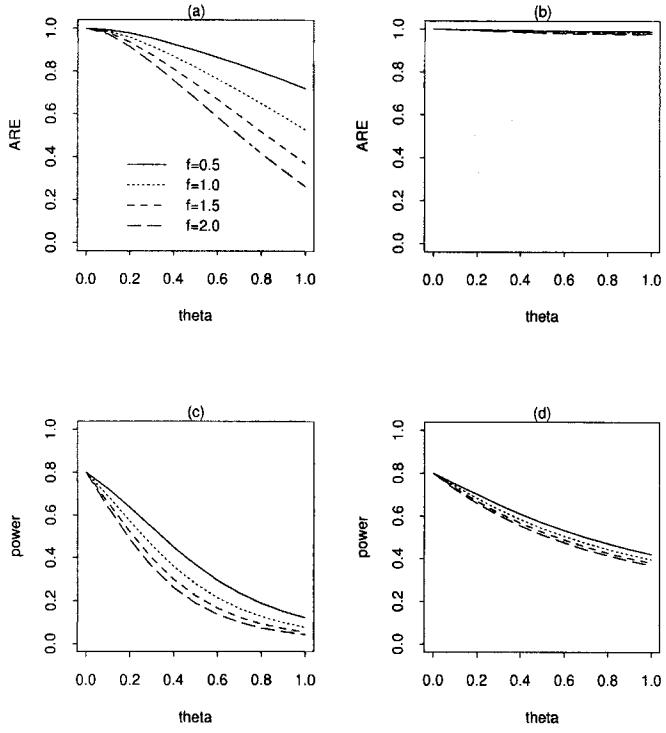

Figure 1. Normal frailty model: θ is the variance of b, the frailty, and $f + 0.5$ is the mean time to censoring.

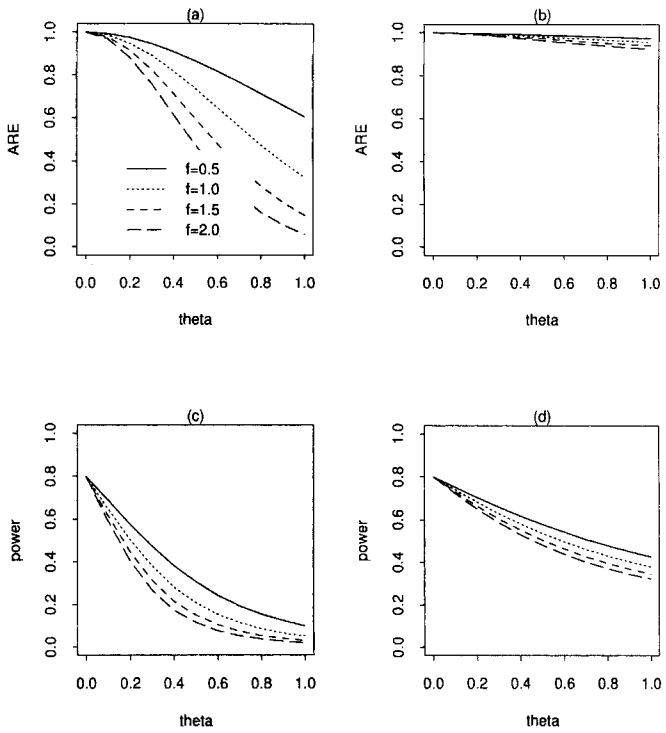

Figure 2. Log-gamma frailty model: θ is the variance of $\exp(b)$, where b is the frailty, and $f + 0.5$ is the mean time to censoring.

6. Numerical Studies

In this section, we evaluate numerically, under several parameter configurations, the ARE of the simple log-rank test relative to the optimally weighted log-rank test and the power of the simple log-rank test for a study designed under the incorrect model (2). We design our numerical calculations in the context of a clinical trial with an accrual period a and a follow-up period f assuming that patients enter the study at a constant rate and are randomized to the two treatments with equal probability. The follow-up period is often introduced in clinical trials to reduce the number of patients needed (i.e., to increase the number of failures observed). Hence, the potential censoring time for patient i is $C_i = a \times U_i + f$, where the $\{U_i, i = 1, \ldots, n\}$ are i.i.d. uniform random variables on the interval $[0, 1]$. This corresponds to a censoring distribution of

$$G(t) = \min\left\{\max\left(1 - \frac{t-f}{a}, 0\right), 1\right\}.$$

In the following calculation, we keep the accrual period fixed at $a = 1$, we let $f = 0.5, 1, 1.5, 2$, and we assume a constant baseline hazard function ($\lambda_0(t) = 1$) in model (1). Finally, we vary the frailty parameter $\theta = 0.0, 0.1, 0.2, \ldots, 1.0$.

We examine the ARE and power under both the no-interaction model (i.e., $\alpha = 0$) and the interaction model (i.e., $\alpha \neq 0$) and for normal and log-gamma frailty distributions, with $b \sim \mathrm{N}(0, \theta)$ and $\exp(b) \sim \mathrm{gamma}(\theta^{-1}, \theta^{-1})$. For the normal frailty model, θ is the variance of b, and for the log-gamma frailty model, θ is the variance of $\exp(b)$. Last, in our evaluations of power, we take $\beta = 0.3$ and $\alpha = 0.3$.

Figure 1 displays the ARE and power as functions of θ for the normal frailty model. For the model with interaction, Figure 1a shows that the ARE decreases with increasing variance of b (i.e., θ). Further, at fixed θ, the ARE decreases with the mean time to censoring; with more censoring, there is less of an opportunity for the heterogeneity to have an impact. In contrast, Figure 1b shows that the ARE remains very close to one over the range of θ and mean times to censoring in the model without interaction. Thus, it appears that, as long as the heterogeneity does not interact with treatment, i.e., patients of different genetic subtypes respond in the same way to the treatment, the simple log-rank test is nearly fully efficient relative to the more complicated optimally weighted log-rank test.

Figure 1c and 1d displays the power as a decreasing function of θ for all different follow-up periods. Also, at fixed θ, it is decreasing in the mean time to censoring. It should be noted that the specified follow-up period corresponds to a specific accrual rate to guarantee a certain power under the naive model. For example, to obtain 80% power under the naive model (2), $f = 0.5, 1, 1.5, 2$ correspond to accrual rates of 900, 660, 567, and 520, respectively. At $\theta = 1$, the power drops to about 15% in the model with interaction and to about 30% in the model without interaction. Figure 1b and 1d shows that, although the simple log-rank test achieves nearly full efficiency relative to the optimally weighted log-rank test, it requires a sufficiently large sample size to achieve 80% power; i.e., if the study is designed assuming the simple hazard model (2), the simple log-rank test may be severely underpowered.

Figure 2 displays the same results for the log-gamma frailty model. The qualitative conclusions from the normal frailty

model hold for this model as well. The one minor difference is that the ARE for the model without interaction drops slightly lower at $\theta = 1$ yet remains above 90%.

7. Discussion

Increasingly, genetic studies of tumors of the same histologic diagnosis are elucidating subtypes that are distinct with respect to clinical endpoints such as response to treatment and survival. This raises concerns about the efficiency of using the simple log-rank test for analysis of treatment effect in studies of possibly heterogeneous tumors. Further, the power of such studies, designed under the assumption of homogeneity, is of serious concern. For these reasons, we undertook this investigation of ARE and power of the simple log-rank test under models of continuous heterogeneity that may interact with treatment.

Based on our numerical studies, we conclude that the simple log-rank test is nearly fully efficient relative to the optimally weighted log-rank test if the unobserved heterogeneity among patients does not interact with treatment; i.e., if the effect of treatment is the same among all patients, regardless of the individual frailty, the simple test is efficient. This is not the case, however, if the frailty interacts with treatment, as appears to be the situation among patients with anaplastic oligodendroglioma. It is now known that patients whose tumors are of one genetic subtype have a much greater response rate to chemotherapy than do patients whose tumors are not of this genetic subtype. Thus, our results suggest that an optimally weighted log-rank test should be used to analyze the on-going clinical trials of anaplastic oligodendroglioma that have not recorded genetic alterations, as the simple log-rank test will suffer from lack of efficiency.

Further, our results suggest that lack of power is a real concern for the on-going clinical trials that were designed assuming the simple model (2) to be true. This is true both when the frailty interacts with treatment and when it does not. Thus, while the simple log-rank test may theoretically achieve a certain level of efficiency, it may also be severely underpowered in practice. This suggests that, if there are suspicions of heterogeneity among the patients, an adaptive design should be implemented to guarantee that the test is fully powered.

Acknowledgements

This research was supported in part by NIH grants CA75971, CA57253, and CA57683.

Résumé

De plus en plus d'études génétiques de tumeurs qui ont le même diagnostic histologique ont fait apparaître des sous-types différents pour la réponse au traitement ou la survie mesurées lors des bilans cliniques intermédiaires. Ceci soulève le problème de l'utilisation du test simple du log-rank pour l'analyse de l'effet du traitement sur la survie dans des études de tumeurs potentiellement hétérogènes. En outre, de telles études, construites sous l'hypothèse d'homogénéité, peuvent avoir une perte sévère de puissance. Nous proposons des approximations analytiques pour l'efficacité relative asymptotique du test simple du log-rank par rapport au test du log-rank pondéré de façon optimale et pour la puissance du test simple du log-rank quand on l'applique à des sujets avec une hétérogénéité non-observée, traduite par une variable de fragilité continue qui peut interagir avec le traitement. Des études numériques montrent que le test simple du log-rank peut être assez inefficace si la fragilité interagit avec le traitement. De plus il peut y avoir une perte substantielle de puissance en présence de fragilité qu'elle interagisse ou non avec le traitement.

References

Carroll, R., Rupert, D., and Stephanski, L. A. (1995). *Measurement Error in Nonlinear Models.* New York: Chapman and Hall.

Chastang, C., Byar, D., and Piantadosi, S. (1988). A quantitative study of the bias in estimating the treatment effect caused by omitting a balanced covariate in survival analysis. *Statistics in Medicine* **7,** 1243–1255.

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.

Keiding, N. and Andersen, P. K. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine* **16,** 215–224.

Lagakos, S. W. and Schoenfeld, D. A. (1984). Properties of proportional-hazards score tests under misspecified regression models. *Biometrics* **40,** 1037–1048.

Morgan, T. M. (1986). Omitting covariates from the proportional hazards model. *Biometrics* **42,** 993–995.

Oakes, D. and Jeong, J. (1998). Frailty models and rank tests. *Lifetime Data Analysis* **4,** 209–228.

Schmoor, C. and Schumacher, M. (1997). Effects of covariate omission and categorization when analysing randomized trials with the Cox model. *Statistics in Medicine* **16,** 225–237.

Schoenfeld, D. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics* **39,** 499–503.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73,** 363–369.

Received August 2001. *Accepted September* 2001.