Journal of Statistical Research *ISSN 0256 - 422 X*
2012, Vol. 46, No. 2, pp. 219-232

MAXIMUM LIKELIHOOD ESTIMATION OF OPTIMAL WEIGHT FUNCTION FOR WEIGHTED LOG-RANK TEST

QING XU

Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD 20993, U.S.A.
Email: xu.qing@fda.hhs.gov

NICHOLAS J. CHRISTIAN AND JONG-HYEON JEONG

Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, U.S.A.
Email: jeong@nsabp.pitt.edu

SUMMARY

We revisit the optimal weights for the weighted log-rank test for nonproportional hazards data. It is noted that the optimal weight function can be derived by assuming a stable distribution for an exponentiated omitting covariate from the proportional hazards model, which induces the nonproportionality. A special case is the weight function for the popular Harrington-Fleming's G^ρ test statistic. However, in practice it is not straightforward for investigators to determine the optimal value of the tuning parameter ρ for the weight function in the G^ρ test statistic. We propose a maximum likelihood method to estimate the parameter from the observed data, noticing that the parameter ρ is inversely related to the index parameter from the gamma distribution commonly assumed for the frailty model. The simulation results indicate that the test statistic with the estimated weight function from the data are more powerful than the commonly used Harrington-Fleming test with $\rho = 1$. We also propose a different weight function that possibly gives more power than existing ones to detect middle difference. Three datasets from phase III clinical trials on breast cancer are illustrated as real examples.

Keywords and phrases: Censoring; Frailty; G^ρ family; Inverse Gaussian; Survival Data.

1 Introduction

The most popular statistical test procedure to compare censored failure time distributions is the simple log-rank statistic (Savage, 1956; Mantel, 1966; Peto, 1972). It is well known, however, that the log-rank test for equality of two failure time distributions is not optimal under nonproportional hazards. The nonproportionality might be caused by omitting a balancing covariate from the proportional hazards model (Lagakos and Schoenfeld, 1984; Morgan, 1986; Struthers and Kalbfleisch, 1986; Oakes and Jeong, 1998), or due to a diminishing treatment effect. In many practical examples from medical research, the hazard rates converge between treatment groups as time progresses. For

example, in a breast cancer study performed by the National Surgical Breast and Bowel Project (NSABP), the effect of tamoxifen, a hormonal therapy, tends to diminish over time, as will be shown in the real data example later.

In this paper, we revisit the optimal weight functions for the weighted log-rank test. We note that the optimal weight function can be derived by assuming a stable distribution for an exponentiated omitting covariate from the proportional hazards model, which induces the nonproportionality (Oakes and Jeong, 1998). Assumption of a gamma distribution for the exponentiated omitting covariate gives the weight function for the popular Harrington-Fleming's G^ρ test. In practice, however, data analysts need to choose a value of ρ arbitrarily, after eyeballing the pattern of difference in hazard or survival rates between groups. For example, the common choice of ρ to test early difference among comparison groups is 1, which will be referred to as Peto-Peto-Prentice test throughout the paper. We propose a maximum likelihood estimation method here to estimate ρ, which turns out to provide higher power than the frequently used Peto-Peto-Prentice test. We also investigate a new weight function derived by assuming an inverse Gaussian distribution for the omitting covariate term, which is also shown to be more powerful than existing methods.

In Section 2, existing linear rank statistics for censored survival data are reviewed. In Section 3, the optimal weight functions, including a new one, for the weighted log-rank test statistic are discussed. In Section 4, the maximum likelihood estimation of the tuning parameter for the weight function is proposed. In Section 5, simulation studies are performed to compare the type I error probabilities and powers of the weighted log-rank test statistic with various weight functions. In Section 6, the proposed test statistic is applied to 3 real datasets from clinical trials on breast cancer. In Section 7, we conclude with a brief remark.

2 Linear Rank Statistics–Review

The simple log-rank test statistic can be derived as a score function from the Cox's proportional hazards model (Cox, 1972). With one binary covariate $x_i = 0$ or 1 for the ith subject, the Cox model specifies

$$h_i(t; x_i) = e^{\beta x_i} h_0(t), \tag{2.1}$$

where $h_0(t)$ is an arbitrary baseline hazard function. Notationally, suppose that survival data consist of independent right censored samples from $K = 2$ populations, and $t_1 < t_2 < \ldots < t_D < \tau$ are the distinct event times in the pooled sample. At time t_i, we observe d_{ij} events in the jth sample among Y_{ij} individuals at risk, $j = 1, 2$, $i = 1, \ldots, D$, $d_i = \sum_{j=1}^2 d_{ij}$ and $Y_i = \sum_{j=1}^2 Y_{ij}$ are the number of events and number of individuals at risk in the combined sample at time t_i. After taking the first derivative of the partial likelihood function (Cox, 1975) with respect to β under model (2.1) and evaluating it at $\beta = 0$ gives the log-rank test statistic to test $H_0 : S_1(t) = S_2(t)$ for all $t \le \tau$ as

$$Z_1(\tau) = \sum_{i=1}^{D} \Big(d_{i1} - Y_{i1} \frac{d_i}{Y_i} \Big), \tag{2.2}$$

so that a general class of weighted log-rank tests takes the form of

$$Z_1^{(W)}(\tau) = \sum_{i=1}^{D} W(t_i)\Big(d_{i1} - Y_{i1}\frac{d_i}{Y_i}\Big). \tag{2.3}$$

Harrington and Fleming (1982) proposed a general class of test statistics (G^ρ test) against specific alternative hypotheses of nonproportional hazard rates by introducing a weight function to the simple log-rank test statistic. The weight function for the G^ρ test is given by

$$W^{(HF)}(t_i) = \{\hat{S}_{pooled}(t_{i-1})\}^\rho, \tag{2.4}$$

where $\hat{S}_{pooled}(t)$ is the Kaplan-Meier product-limit estimator of the survival function based on the pooled data. The simple log-rank test (Peto and Peto, 1972) is obtained when $\rho = 0$. A case of $\rho = 1$ gives a modified version of the Wilcoxon test, i.e. the Peto-Peto-Prentice test. When $\rho > 0$, the weight function gives more weights to early differences between the hazard rates. On the contrary, when $\rho < 0$, it assigns more weights on late differences.

3 Optimal Weight Functions

Oakes and Jeong (1998) derived a new class of optimal weight functions for the weighted log-rank test for nonproportional hazards data. They considered a case where a covariate that balances the proportionality in the Cox model has been omitted, introducing nonproportionality. They have derived various weight functions explicitly by assuming some parametric distributions for the exponentiated term of the omitted covariate, adopting the frailty theory (Vaupel *et al.*, 1979).

For the ith subject, the frailty model specifies

$$h_i(t \mid x_i, z_i) = v_i e^{\beta x_i} h_0(t), \tag{3.1}$$

where $h_0(t)$ is an unknown baseline hazard function, x_i is a binary covariate, for simplicity, and $v_i = e^{\nu z_i}$ is a frailty. In general, the frailty may imply an unobservable genetic or environmental heterogeneity in the population, so that a distribution needs to be assumed for v_i. Note that the model (3.1) satisfies the assumption of proportional hazards given x_i and z_i, implying that omitting the second covariate z_i, possibly continuous, might result in nonproportionality. By using the survival function, the model (3.1) can be written as

$$S_i(t; x_i, z_i) = \exp(-B v_i \theta_i), \tag{3.2}$$

where $B = B(t) = -\log S_0(t)$, $S_0(t)$ being the baseline survival function, and $\theta_i = \exp(\beta x_i)$. Now by assuming that the model (3.2) is the true model for the data at hand but the term v_i was not observed or omitted, we can write a marginal survival function

$$S(t; x_i) = E_V\{\exp(-B\theta_i V)\} = p(\theta_i B), \tag{3.3}$$

where $p(\cdot)$ is the Laplace transform for the distribution of the frailty V. Various distributions have been proposed for V such as gamma, Inverse Gaussian, or positive stable (Hougaard, 1984, 1986).

When the frailty variable V follows a gamma distribution with both mean and variance of κ, the marginal survival function (3.3) reduces to

$$S_G(t;\theta_i) = \left(\frac{1}{1+\theta_i B}\right)^{\kappa}, \tag{3.4}$$

which gives the hazard ratio as

$$\frac{e^{\beta}(1+B)}{1+e^{\beta}B}.$$

Note that the hazard ratio between the two groups are not proportional, changing from e^{β} to 1 as $t \to \infty$.

A popular form of the inverse Gaussian frailty distribution with the unit mean and variance of $(2\phi)^{-1}$ is given by (Hougaard, 1984)

$$S_{IG}(t;\theta_i) = \exp\left[2\phi - 2\sqrt{\phi(\phi+\theta_i B)}\right]. \tag{3.5}$$

In this case, the hazard ratio is

$$e^{\beta}\sqrt{\frac{\phi+B}{\phi+e^{\beta}B}}.$$

Again the hazard ratio is not proportional, changing from e^{β} to $\sqrt{e^{\beta}}$ as $t \to \infty$.

Oakes and Jeong (1998) showed that under the null hypothesis of $H_0 : \beta = 0$, the optimal weight function $W(t)$ in the weighted log-rank test (2.3) converges to

$$w(t) = 1 + \frac{Bp''(B)}{p'(B)} - \frac{Bp'(B)}{p(B)}. \tag{3.6}$$

Under the gamma frailty assumption, the equation (3.6) gives

$$w_G(t) = \frac{1}{1+B} = \left[\left(\frac{1}{1+B}\right)^{\kappa}\right]^{1/\kappa} = S_G(t)^{\rho}, \tag{3.7}$$

where $\rho = 1/\kappa$, and $S_G(t) = S_G(t;\theta_i)|_{\theta_i=1}$ from (3.4). As noted in Oakes and Jeong (1998), this implies that the optimal weight function derived under the assumption that the exponentiated omitting covariate follows a gamma distribution is equivalent to one for the Harrington and Fleming's G^{ρ} test statistic. Note that $w_G(t)$ tends to 1 as $\kappa \to \infty$, which leads to the simple log-rank test. Similarly, under the inverse Gaussian frailty assumption, the optimal weight function can be derived as

$$w_{IG}(t) = 1 - \frac{B}{2(\phi+B)} = \frac{1}{2} + \frac{2\phi^2}{\{2\phi - \log S_{IG}(t)\}^2}, \tag{3.8}$$

where $S_{IG}(t) = S_{IG}(t;\theta_i)|_{\theta_i=1}$ from (3.5).

Hougaard (1984) showed that, relative to the gamma distribution, under the inverse Gaussian frailty model the surviving population becomes more homogeneous with time. This might imply that the weighted log-rank test with the new weight function in (3.8) could be more powerful when a group effect fades away over time, two groups being homogeneous.

4 Estimation of a Tuning Parameter for the Weight Function

Even though the G^ρ family is rich, encompassing a variety of weighting schemes, in practice it is not easy for investigators to determine which value for the tuning parameter ρ should be used, so that conveniently the test statistic such as Peto-Peto-Prentice test has been often adopted, fixing the parameter value as 1, to infer early difference in survival data. However, the equation (3.7) suggests that the pattern of nonproportionality caused by an omitted covariate information following a gamma distribution can be explained by the weight function from the G^ρ family, in which case $\rho = 1/\kappa$. Hence the tuning parameter ρ can be estimated from the model given in (3.4), assuming a parametric form for the baseline cumulative hazard function B. For the i^{th} subject, let $\delta_i = 1$ if an event occurred at time T_i and $\delta_i = 0$ if an observation was censored at time T_i. Then for n observations under right censoring, the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} h_G(T_i;\theta_i)^{\delta_i} S_G(T_i;\theta_i)$$

where $S_G(t;\theta_i)$ is the marginal survival function (3.4) and $h_G(t;\theta_i) = -\frac{d}{dt}\log(S_G(t;\theta_i))$.

Specifically, let us assume an exponential distribution with mean λ for the baseline hazard function and a gamma frailty distribution with both mean and variance of κ. Then the likelihood function is given by

$$L(\kappa,\beta,\lambda) = \prod_{i=1}^{n} \left(\frac{\lambda\kappa e^{\beta x_i}}{1+\lambda t_i e^{\beta x_i}}\right)^{\delta_i} \left(\frac{1}{1+\lambda t_i e^{\beta x_i}}\right)^{\kappa}, \tag{4.1}$$

which can be maximized using an optimization procedure. Applying the invariance property of the maximum likelihood estimator (MLE), the estimate of κ, $\hat{\kappa}$, can be used to estimate ρ in the gamma frailty weight function (3.7). Similar steps can be applied to estimate the parameter ϕ for the inverse Gaussian frailty function in (3.5).

5 A Simulation Study

In this section, first we compare the type I error probabilities and powers of the weighted log-rank test statistic using different weight functions when (3.4) is the true model. We will consider three weight functions; simple log-rank ($W(t) = 1$), Fleming-Harrington with $\rho = 1$ ($W(t) = \hat{S}_{pooled}(t_{i-1})$) and the gamma frailty weight function ($W(t) = \hat{S}_{pooled}(t_{i-1})^{\hat{\rho}}$), where $\hat{\rho}$ is the maximum likelihood estimate of ρ.

In the simulation study, we assume an exponential baseline hazard with mean 2. The covariate values of x's were generated from a Bernoulli distribution with mean 0.5. Event times conditional on x's were generated from (3.4) using the probability integral transformation. Censoring times were generated independently from a uniform 0 to c distribution, where c is chosen to achieve the desired censoring proportion. Simulations were performed for a sample size of $n = 300$ and censoring proportions of 0%, 30% and 60%. The true values were 0.1, 0.25, 0.5, 1 for κ and 0, 0.5 and 0.75 for β.

One thousand samples were drawn from each configuration of β, κ, and the censoring proportion. For each sample, a test for the null hypothesis of no difference between the two failure distributions was conducted using each of the three weighted log-rank tests described above. Then the proportion of statistically significant cases at the significance level of 0.05, was obtained for each test. The maximum likelihood estimate of κ was also calculated for each sample, using `R`'s `constrOptim` function. This function maximizes the likelihood function using an adaptive barrier algorithm with linear inequality constraints on the parameters. In this case κ and λ were constrained to be greater than or equal to 0, but β was not.

Table 1 summarizes the true values and the mean of the maximum likelihood estimates of κ, and the mean square errors of the estimates for various censoring proportions when $\beta = 0.75$. The result indicates that the estimates are very close to their true values on average except the unit exponential frailty case, i.e. when $\kappa = 1$, with heavy censoring.

Table 2 summarizes type I error probabilities ($\beta = 0$) and powers ($\beta = 0.75$) for the three weighted log-rank tests for various values of κ and different censoring proportions; (1) weight function from the gamma frailty model with $\hat{\rho} = 1/\hat{\kappa}$ (Gamma Frailty), (2) simple log-rank test, and (3) Fleming-Harrington's G^ρ test with ρ fixed as 1. All three tests give reasonable type I error probabilities. In terms of powers, the simple log-rank test produced the lowest powers, as expected, and the G^ρ statistic with $\hat{\kappa}$ provided the highest powers. As the true values of κ get closer to 1, the results from the tests with gamma frailty weighting function and the G^ρ weighting function with $\rho = 1$ converges.

The simulation was repeated for $\beta = 0.5$, and both results are graphically presented in Figure 1, which indicates that the efficiency of the G^ρ test with the estimated weight function is higher for data with lower censoring proportion and small/middle values of κ.

The similar simulation was performed assuming an inverse Gaussian frailty distribution in (3.3), and the results are summarized in Table 3, where the tuning parameter is ϕ for the inverse Gaussian weighting function. In this case, the test with the estimated inverse Gaussian weight function performs slightly better in case of no/moderate censoring, and the results are compatible with the fixed G^ρ test and simple log-rank test as the censoring proportion increases, due to the heavy censoring effect at the tail.

For a fair comparison, we have also performed power analysis for the G^ρ test statistic when the true distribution was misspecified as the inverse Gaussian distribution. Table 4 compares the estimated G^ρ test statistic, simple log-rank test statistic, and Fleming-Harrington test with $\rho = 1$. The results indicate that the estimated G^ρ test and log-rank test tends to perform slightly better than the fixed G^ρ test with $\rho = 1$.

6 Real Data Examples

In this section, three real datasets from phase III clinical trials on breast cancer, performed by National Surgical Breast and Bowel Project (NSABP), are illustrated to compare the three types of weighted log-rank statistics.

The first dataset is from a study (B-13) that assessed sequential Methotrexate $\rightarrow$ 5-Fluorouracil

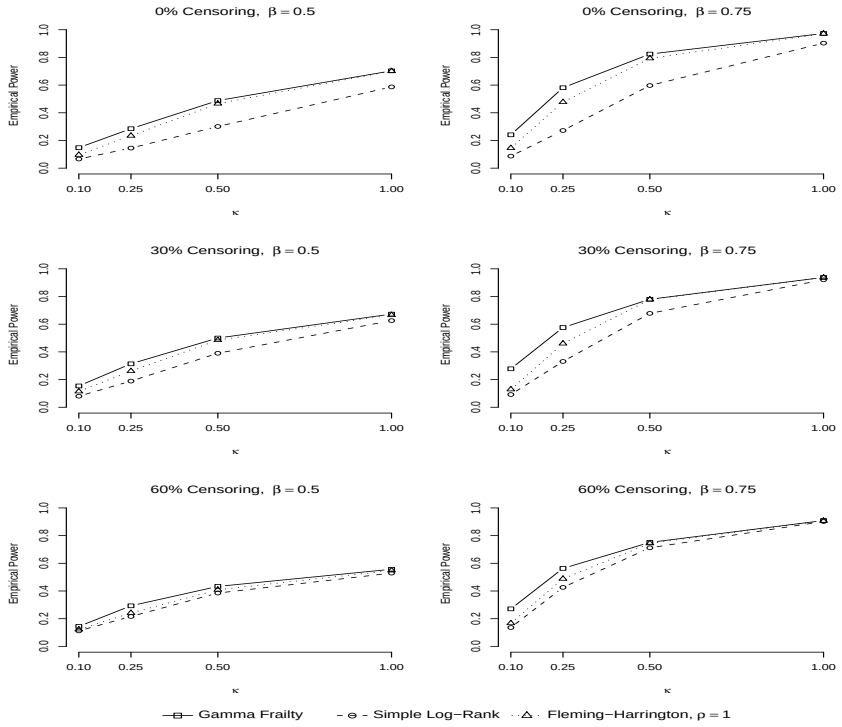

Figure 1: Power comparisons between simple log-rank test statistic, G^ρ test with the estimated weight function (Gamma Frailty), and G^ρ test with fixed $\rho = 1$ (Fleming-Harrington, $\rho = 1$); censoring proportion= 0%, 30%, and 60%; κ=0.1, 0.25, 0.50, and 1.0; $\beta = 0.5$ and 0.75.

(M→F) in breast cancer patients with negative axillary lymph nodes and negative estrogen receptors. The main results have been published and updated previously (Fisher *et al.*, 1989b, 1996b, 2004). In the analysis presented here included were total 731 eligible patients with follow-up information (369 in placebo group; 362 in the M→F group). The second dataset comes from the NSABP B-14 study, where patients with primary breast cancer, negative axillary nodes, and estrogen receptor positive tumors were randomized to receive either tamoxifen (a hormonal therapy) or placebo following surgery. The trial itself is described in details in the literature (Fisher *et al.*, 1989a, 1996a). Only total 68 eligible patients with tumor size greater than 5cm (30 from placebo group; 38 from tamoxifen group) were used in this analysis. The third dataset comes from NSABP B-19 protocol, where the similar patient population (node-negative and ER-negative) has been studied as in B-13 to compare the M→F regimen with the conventional Cyclophosphamide, Methotrexate, and 5-Fluorouracil (CMF) regimen. A cohort of 1,074 eligible patients have been included from B-19 study. The mean time on study ranges from about 17 years to 21 years. In these studies, the endpoints of interest were disease-free survival (DFS) and overall survival (OS). The DFS endpoint

includes breast cancer recurrences, other primary cancers, and deaths as first events, and the OS endpoint includes any deaths. The censoring proportions for B-13, B-14, and B-19 data were 68%, 47%, and 76%, respectively.

Figure 2 shows the smoothed hazard plots between comparison groups of the OS endpoint (B-13, and B-19) and the DFS endpoint (B-14). The early difference is more noticeable in the B-14 data, but the hazard rates converge over time in all three datasets. P-values from the simple log-rank test were 0.0204, 0.411, and 0.011 for B-13, B-14, and B-19 data, respectively. P-values from the G^ρ test with $\rho = 1$ were 0.019, 0.269, and 0.0097. P-values from the G^ρ test with the estimates $\hat{\rho} = 4.93$, 3.22, and 2.71 were 0.0235, 0.170, and 0.0077, indicating that the G^ρ test with the weight function estimated by the maximum likelihood provides more efficient results for B-14 and B-19 data. For B-13 data, the results were compatible between the Peto-Peto-Prentice test and the estimated G^ρ test. This phenomenon was also observed in the simulation results presented in Figure 1; the efficiency of the test based on the maximum likelihood estimation of the tuning parameter tends to be higher for data with lower censoring proportion and large/middle value of $\rho = 1/\kappa$. The test with the estimated inverse Gaussian weighting function gives p-values of 0.022, 0.385, and 0.0091, respectively, indicating that it performs as well as the previous two G^ρ statistics for the B-13 and B-19 data, but not for B-14 data. Note that the inverse Gaussian weight function assigns more weights for the middle difference, compared to the gamma weight function.

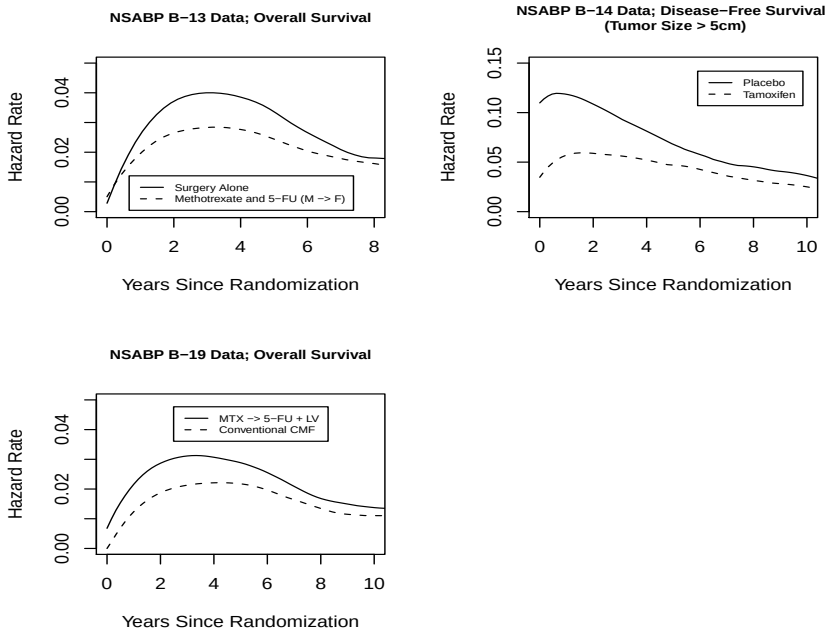

Figure 2: Smoothed hazard rates between treatment groups (NSABP B-13, B-14, and B-19 data).

7 A Concluding Remark

In this paper, we considered a new weighting function for the weighted log-rank test and estimation of the tuning parameter for the weight function, when the hazard rates converge as time progresses. The pattern of weighting under the Harrington-Fleming's G^ρ statistic can be captured by the optimal weight function derived from the gamma frailty model, indicating a connection between the parameter ρ and the parameter from the assumed gamma frailty. Therefore the tuning parameter could be estimated from the marginal distribution induced under the gamma frailty model. The simulation results and the real data examples indicate that the maximum likelihood estimates of the tuning parameter for the weight function and the estimated G^ρ test performs better than one with an arbitrarily fixed value, such as the Peto-Peto-Prentice test. In our simulation and real data examples, however, we assumed an exponential distribution for the baseline hazard distribution for the maximum likelihood estimation. Even though the exponential distribution would capture the overall pattern of hazard rate reasonably well on average, a nonparametric approach would allow for more flexibility in practice, which is under investigation.

Acknowledgement

We thank an anonymous reviewer for his/her helpful comments. This research was supported in part by the Department of Defense (DOD) grant W81XWH-04-1-0605, and National Health Institute (NIH) grants 5-U10-CA69974-09 and 5-U10-CA69651-11.

References

[1] Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society B* **34**, 187-220.

[2] Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269-276.

[3] Fisher, B., Costantino, J., Redmond, C. *et al.* (1989a). A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors. *New England Journal of Medicine* **320**, 479-484.

[4] Fisher, B., Dignam, J., Bryant, J. *et al.* (1996a). Five versus more than five years of tamoxifen therapy for breast cancer patients with negative lymph nodes and estrogen receptor-positive tumors. *Journal of the National Cancer Institute* **88**, 1529-1542.

[5] Fisher, B., Dignam, J., Mamounas, T., *et al.* (1996b). Sequential methotrexate and 5-fluorouracil (M→F) in the treatment of node-negative breast cancer patients with estrogen receptor-negative tumors: eight-year results from NSABP B-13 and first report of findings from NSABP B-19 comparing M→F with conventional CMF. *Journal of Clinical Oncology* **14**, 1982-1992.

[6] Fisher, B., Jeong, J., Anderson, S., *et al.* (2004). Treatment of axillary lymph node-negative, estrogen receptor-negative breast cancer: updated findings from National Surgical Adjuvant Breast and Bowel Project clinical trials. *Journal of National Cancer Institute* **96**, 1823-1831.

[7] Fisher, B., Redmond, C., Dimitrov, N., *et al.* (1989b). A randomized clinical trial evaluating sequential methotrexate and fluorouracil in the treatment of patients with node-negative breast cancer who have estrogen receptor-negative tumors. *New England Journal of Medicine* **320**, 473-478.

[8] Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69**, 553-566.

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

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

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

[12] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50**, 163-170.

[13] Morgan, M. M. (1986). Omitting covariates from the proportional hazards model. *Biometrics* **42**, 993-995.

[14] Oakes, D. and Jeong, J.-H. (1998). Frailty model and rank tests. *Lifetime Data Analysis* **4**, 209-228.

[15] Peto, R. (1972). Contribution to the discussion of a paper by D. R. Cox. *Journal of the Royal Statistical Society, Series B* **34**, 205-207.

[16] Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant procedures. *Journal of the Royal Statistical Society, Series A* **35**, 185-207.

[17] Savage, I. R. (1956). Contributions to the theory of rank order statistics-the two-sample case. *Annals of Mathematical Statistics* **27**, 590-615.

[18] Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazards models. *Biometrika* **73**, 363-369.

[19] Vaupel, J. A., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439-454.

Table 1: True values and maximum likelihood estimates with mean square errors (MSE) of κ; censoring proportion=0%, 30%, and 60%; $\beta = 0.75$.

Censoring Rate	κ	$\hat{\kappa}$	MSE($\hat{\kappa}$)
0%	0.10	0.100	0.00004
	0.25	0.252	0.00033
	0.50	0.507	0.00229
	1.00	1.020	0.01483
30%	0.10	0.100	0.00006
	0.25	0.252	0.00063
	0.50	0.510	0.00508
	1.00	1.044	0.05044
60%	0.10	0.101	0.00015
	0.25	0.261	0.00239
	0.50	0.538	0.02839
	1.00	1.295	0.98234

Table 2: Empirical type I error probabilities ($\beta = 0$) and powers ($\beta = 0.75$) from simple log-rank test statistic, G^ρ test with the estimated gamma frailty weight function (Gamma Frailty), and G^ρ test with fixed $\rho = 1$ (Fleming-Harrington, $\rho = 1$); censoring proportion= 0%, 30%, and 60%; κ=0.1, 0.25, 0.50, and 1.0.

	Censoring Rate	κ	Gamma Frailty	Simple Log-Rank	Fleming-Harrington, $\rho = 1$
$\beta = 0$	0%	0.10	0.038	0.057	0.053
		0.25	0.044	0.057	0.058
		0.50	0.050	0.051	0.045
		1.00	0.048	0.048	0.048
	30%	0.10	0.056	0.043	0.044
		0.25	0.052	0.062	0.061
		0.50	0.043	0.041	0.045
		1.00	0.058	0.064	0.059
	60%	0.10	0.051	0.039	0.039
		0.25	0.050	0.063	0.063
		0.50	0.068	0.066	0.064
		1.00	0.052	0.047	0.051
$\beta = 0.75$	0%	0.10	0.241	0.087	0.145
		0.25	0.581	0.272	0.477
		0.50	0.824	0.597	0.793
		1.00	0.973	0.903	0.971
	30%	0.10	0.278	0.092	0.129
		0.25	0.576	0.331	0.460
		0.50	0.780	0.678	0.778
		1.00	0.938	0.921	0.937
	60%	0.10	0.271	0.136	0.167
		0.25	0.563	0.426	0.488
		0.50	0.751	0.712	0.744
		1.00	0.909	0.902	0.907

Table 3: Empirical type I error probabilities ($\beta = 0$) and powers ($\beta = 0.5$) from simple log-rank test statistic, weighted linear rank test with the estimated inverse Gaussian frailty weight function (Inverse Gaussian Frailty), and G^{ρ} test with fixed $\rho = 1$ (Fleming-Harrington, $\rho = 1$); censoring proportion= 0%, 30%, and 60%; κ=0.1, 0.25, 0.50, and 1.0.

	Censoring Rate	ϕ	Inverse Gaussian Frailty	Simple Log-Rank	Fleming-Harrington, $\rho = 1$
$\beta = 0$	0%	0.10	0.055	0.055	0.055
		0.25	0.042	0.040	0.034
		0.50	0.051	0.047	0.049
		1.00	0.048	0.049	0.055
	30%	0.10	0.049	0.045	0.049
		0.25	0.047	0.050	0.052
		0.50	0.058	0.053	0.056
		1.00	0.049	0.053	0.056
	60%	0.10	0.054	0.049	0.058
		0.25	0.059	0.050	0.059
		0.50	0.050	0.045	0.051
		1.00	0.047	0.043	0.048
$\beta = 0.5$	0%	0.10	0.726	0.715	0.659
		0.25	0.798	0.797	0.748
		0.50	0.868	0.861	0.831
		1.00	0.924	0.918	0.881
	30%	0.10	0.590	0.572	0.568
		0.25	0.715	0.696	0.703
		0.50	0.752	0.755	0.742
		1.00	0.848	0.842	0.821
	60%	0.10	0.435	0.435	0.434
		0.25	0.522	0.516	0.520
		0.50	0.609	0.600	0.611
		1.00	0.682	0.686	0.667

Table 4: Empirical power assuming a gamma frailty distribution when the true distribution is an inverse Gaussian frailty distribution; $n = 300$ and $\beta = 0.75$.

Censoring Rate	ϕ	Gamma Frailty	Simple Log-Rank	Fleming-Harrington, $\rho = 1$
0%	0.10	0.649	0.695	0.641
	0.25	0.768	0.794	0.733
	0.50	0.868	0.864	0.833
	1.00	0.933	0.932	0.888
30%	0.10	0.556	0.582	0.566
	0.25	0.710	0.710	0.693
	0.50	0.805	0.791	0.784
	1.00	0.858	0.839	0.835
60%	0.10	0.427	0.429	0.433
	0.25	0.522	0.518	0.524
	0.50	0.620	0.621	0.616
	1.00	0.678	0.674	0.671